March 2019

Visual Recollection for Non-Declarative Representations

Patrick Sadil

VISUAL RECOLLECTION FOR NON-DECLARATIVE REPRESENTATIONS

A Thesis Presented

by

PATRICK S. SADIL

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

February 2019

Cognition and Cognitive Neuroscience

VISUAL RECOLLECTION FOR NON-DECLARATIVE REPRESENTATIONS

A Thesis Presented

by

PATRICK S. SADIL

Approved as to style and content by:

_____
Rosemary A. Cowell, Chair

_____
David E. Huber, Member

_____
Joonkoo Park, Member

_____
Caren Rotello, Department Chair
Psychological and Brain Sciences

ABSTRACT

VISUAL RECOLLECTION FOR NON-DECLARATIVE REPRESENTATIONS

FEBRUARY 2019

PATRICK S. SADIL, B.A., REED COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS, AMHERST

Directed by: Assistant Professor Rosemary A. Cowell

Recollection is a pattern completion process that enables retrieval of arbitrarily associated information following minimal study. These attributes enable recollection to support retrieval of many kinds of mnemonic representations, from highly associative contextual information to very specific low-level representations. However, recollection is typically studied in the context of declarative memory tasks, in which participants exhibit recollection by explicitly reporting on the recollected information. Is it the case that recollection is limited to declarable representations, or is it a more general process that occurs for any representation? Two experiments and a novel analysis technique are presented to answer this question. The results suggest that recollection is not limited to declarable representations. These results argue against theories of recognition memory that restrict the representational input allowed to mnemonic processes; mnemonic processes in general may act on arbitrary representations.

CONTENTS

Page

ABSTRACT....v

LIST OF FIGURES viii

CHAPTER

1. VISUAL RECOLLECTION FOR NON-DECLARATIVE REPRESENTATIONS 1

1.1 Introduction....1

2. CFS STATE TRACE....7

2.1 Introduction....7
2.2 Methods....11

2.2.1 Participants....11
2.2.2 Materials 11
2.2.3 Procedure 12

2.2.3.1 Study Phase....13
2.2.3.2 Test Phase 15

2.2.4 Analyses....16

2.2.4.1 Overview of State-Trace Analysis....16
2.2.4.2 Correction for Guessing....19

2.2.5 Results....21
2.2.6 Discussion....23

3. HIERARCHICAL BAYESIAN STATE-TRACE ANALYSIS....25

3.1 Introduction....25
3.2 Detail of State-Trace Analysis....25
3.3 Hierarchical Multivariate Probit Model for State-Trace Analysis....29
3.4 Model Recovery Via Simulation 36
3.5 State-Trace Analysis Applied to Experiment 1 38
3.6 Remaining Shortcomings....39
3.6 Discussion....40

4. VISUAL RECOLLECTION....42

4.1 Introduction....42
4.2 Methods....46

4.2.1 Participants....46
4.2.2 Procedure....47

4.2.2.1 Ocular Dominance....47
4.2.2.2 Opacity Calibration....47
4.2.2.3 Main Experiment....49

4.3 Analyses....50
4.4 Results....52
4.5 Discussion....54

5. GENERAL DISCUSSION....56

6. FIGURES....61

APPENDICIES

A. METHODS TO ASSESS CONVERGENCE OF HAMILTONIAN

MONTE-CARLO....74

B. WIDELY APPLICABLE INFORMATION CRITERION....77

REFERENCES....82

LIST OF FIGURES

Figure Page

Figure 1 A subset of the representational hierarchy 61

Figure 2: Example stimuli used in the Gollin Incomplete Pictures task. Figure reproduced from Hirshman (1990). 62

Figure 3 Sample test questions in Experiment 1.....63

Figure 4 Crossover interaction was observed in Experiment 1. 64

Figure 5 State-Trace analysis of Experiment 1.....65

Figure 6. Hypothetical cognitive models comparable with state-trace analyses. 66

Figure 7. Negative effect of measurement dependence on STA. 67

Figure 8. Schematic of the Bayesian hierarchical multivariate probit model.....68

Figure 9 Simulated data for STA.....69

Figure 10 Experiment 2 task.....70

Figure 11 Descriptive plots of Experiment 2 performance.....71

Figure 12 Quintile analysis of Experiment 2 72

Figure 13 Posterior of the effect of study on the shift in the RT distribution.....73

CHAPTER 1

VISUAL RECOLLECTION FOR NON-DECLARATIVE REPRESENTATIONS

1.1 Introduction

Memory has classically been divided into more basic processes such as priming, reflex, and fear-conditioning from the more complex including episodic memory and semantic memory. Central to this division is the notion of non-declarative and declarative memory processes, whereby the more basic are non-declarative whereas the others are declarative (Squire & Dede, 2015; Squire & Zola-Morgan, 1991). This thesis focuses on the mnemonic process of recollection. Recollection will be operationalized as pattern-*completion*, whereby an individual is presented with only part of a previously encountered stimulus and must generate details of the rest of that stimulus (Montaldi & Mayes, 2010). Recollection juxtaposes familiarity, which is a pattern-*matching* process (Norman, 2010). That is, familiarity does not generate information about a stimulus but instead outputs an index of how well a stimulus matches stimuli that have previously been encountered. There is much debate about the mechanisms underlying these processes (i.e., threshold vs. continuous-valued; e.g., Greve, Donaldson, & Van Rossum, 2010; Pazzaglia, Dube, & Rotello, 2013; Yonelinas, 2002; Yonelinas & Parks, 2007). However, this distinction between a generative (recollection) and a matching (familiarity) process is generally accepted by dual-process theories of recognition memory (e.g., Norman, 2010; Norman & O'Reilly, 2003).

The processes of recollection and familiarity have been characterized as being differently useful for distinct kinds of memories (Yonelinas, 1999, 2013). Recollection is often characterized as being able to produce arbitrary associations, learned after only

minimal (i.e., "one-shot") study, and the content is often episodic in nature. That is, the representational domain of recollection tends to be linked to a particular time and a particular place. Mandler's classic example of retrieving the details of your butcher and his shop (i.e., complex, contextual information) after encountering him on the bus exemplifies this relationship between recollection and higher-level representations (Mandler, 1980). Likewise, studies that rely on source recognition construct tasks where participants must recollect the context (e.g., gender of speaker, scene image, font color) in which an item (e.g., word, picture of a face) was studied. Recollection is assumed to occur when participants retrieve episodic information. In this thesis, representations that are episodic in nature will be contrasted with information that is learned over longer periods of time – gestalt information about how visual information tends to be organized (where 'tends to be' is something that must be learned through experience), the statistical properties of the visual world[1].

However, the extent to which recollection only occurs for certain kinds of representations is unknown. A restriction that recollection only generates declarative information contrasts with the predictions of an account of memory called the Representational-Hierarchical (RH) account (Cowell, Bussey, & Saksida, 2010; Kent, Hvoslef-Eide, Saksida, & Bussey, 2016; Murray & Bussey, 1999; Saksida & Bussey, 1998). A central claim of the RH account is that mnemonic processes such as recollection

[1] The distinction between episodic and statistical information is somewhat oversimplified. There are instances in which researchers (who claim that recollection and familiarity are differently useful for different kinds of representations) note that familiarity can support the rapid learning of novel associations and content (Mayes et al., 2004). One mechanism by which familiarity can support such learning has been termed 'unitization,' which is predicted to occur when the association to be learned can be processed as a conjunctive whole (as compared to a collection of associated parts) such as when learning faces. Alternatively, familiarity may support the learning of representations within a representational domain (e.g., associations between faces), but not between representational domains (e.g., a face and an object).

are performed on any mnemonic representation. That is, the RH accounts claims that there is not necessarily a reason to hypothesize that pattern-completion can only occur for certain kinds of patterns. The primary question of this thesis will be whether recollection can occur for non-episodic representations. Recollection for non-episodic representations will be called "visual recollection." Visual recollection corresponds to a process whereby, e.g., a participant is given only part of an image of an object (e.g., a picture that shows just the handle of a teapot) as a cue and recalls other, low-level visual details of that object – without necessarily retrieving higher-level representations of the object. Visual recollection will be contrast with "episodic recollection." Episodic recollection is defined as another recollective process (perhaps initiated by the same parts of objects), but one in which higher-level representations are recollected. Episodic and visual recollection are therefore differentiated only by the kinds of representations that they are defined to generate[2]. That is, they are not meant to imply qualitatively different processes or neurocomputational mechanisms. Instead, they are meant to be suggestive of a single pattern-completion process that happens for two different points on a continuum of differently complex representations which range from low-level visual details to highly associative episodic representations. The central question of this thesis is: to what extent can visual recollection proceed without episodic recollection (Figure 1)?

[2] The operational details for how to infer presence of recollection (either visual or episodic) will necessarily depend on the experiment. For example, as discussed above, asking whether participants recollect the source of a studied item is one typical way of measuring episodic recollection. A challenge to studying visual recollection is the need to use a task that enables participants to demonstrate retrieval of a memory, but that doesn't rely on their ability to explicitly declare the retrieved memory. See Chapter 4 for the operational definition of visual recollection utilized in this thesis.

Studies involving individuals with medial-temporal lobe (MTL) damage-induced amnesia have provided some evidence that visual recollection can occur, even when participants have limited episodic representations. For example, in the fragment-completion (also known as Gollin Incomplete Pictures; Figure 2) task, participants are presented with line drawings of objects at progressive stages of completeness, one at a time, from least complete to most complete. Across sessions of being asked to name the same objects, participants learn to identify the objects at earlier stages of completeness. Individuals with MTL-damage induced amnesia are able to improve on this task, despite a lack of an explicit memory earlier sessions (Milner, Corkin, & Teuber, 1968; Warrington & Weixkrantz, 1968). That is, even though such participants have only the low-level cues of the incomplete pictures at the time of test (given that they have no memory of performing a similar task at previous occasions), they are able to learn associations between the fragmented image and the whole object, then retrieve those associations to identify the object from the fragments.

Similar to the performance of individuals with amnesia in the incomplete pictures task, Graf, Squire, & Mandler (1984) famously discovered that individuals with MTL damage-induced amnesia fail to recall previously studied words, but nonetheless complete word-stems with those studied words. Specifically, these individuals failed to generate studied words when they were given the word-stems of those studied words and instructed to list recall the full word that they studied. However, when simply instructed to complete the word stems (i.e., when not instructed to produce words from the study list but instead produce whichever words came to mind), individuals with amnesia completed the stems with words that they studied. That is, even though the cues for these two tasks

were the same (a word-stem of a studied word), individuals with amnesia only generated information about the studied items when given instructions that did not emphasize effortful recall.

Although the successful retrieval of information by individuals with amnesia that are cued by low-level information is suggestive of a visual-recollective process (recollection for non-episodic representation), these studies do not provide unambiguous evidence that visual recollection can proceed without mediation by higher-level representations. Whether the *learning* of intra-item visual associations can proceed without such mediation is not answered by these studies for two reasons: 1) participants were free to associate the low-level details with higher-level representations at the time of study, and 2) the measure of memory was the production of the studied item's name. The first reason makes it unclear whether the retrieval processes utilized by participants with amnesia was contingent on the encoding of associations between lower- and higher-level representations. For example, in the incomplete pictures task, participants may have been able to identify the patterns at an earlier level of incompleteness – not because they had learned "lateral" associations between the fragments – but rather because they had associated the fragmented image with the whole object. Any retrieval of missing visual details might have occurred via a "top-down" association between the whole object and the missing visual details. In this way, any visual recollection that might have occurred would have been mediated by associations between the lower- and higher-level representation.

The second reason highlights a challenge of studying recollection for visual details in a behavioral setting; tasks that measure recollection by asking participants to

explicitly generate at test the names of studied material require participants to produce multiple levels of information. Visual recollection is defined as the pattern-completion of low-level visual details. In this thesis, it is assumed that the semantic identity is a higher-level representation than just the visual details. This assumption follows from the intrinsic link between the semantic identity of an object and its visual conjunctive representation; awareness of the conjunctive visual representations is assumed to occur if and only if there is awareness of the semantic identity of an object. With this assumption, data that attempt to measure visual recollection for low-level visual details are ambiguous as to whether lower-level representations could be generated without simultaneous generation of higher-level representations. Considering again the fragment completion task, it is seen that an inability to identify an object from a largely incomplete image of that object does not necessarily measure the extent to which some parts of that object are still retrieved.

This thesis presents two experiments and a novel analysis technique to ask whether visual recollection proceed for lower-level visual representations without mediation by higher-level representations. The experiments were designed to bypass the two challenges listed above. The first experiment asks whether the learning of lateral associations between the visual details of objects can be decoupled from the learning of associations with higher-level representations. A novel analysis technique that builds on state-trace analyses will be developed to handle the data produced by the first experiment. The second experiment asks whether lateral associations between lower-level representations can support pattern-completion of lower-level visual representations.

CHAPTER 2

CFS STATE TRACE

2.1 Introduction

The first experiment examined whether the learning of lateral associations between the visual details of objects can be decoupled from the learning of associations with higher-level representations. Lower-level representations were operationalized as intra-item parts of objects, and higher-level representations were operationalized as the identity of the object (Figure 1). In this experiment, participants studied objects in two conditions that were hypothesized to differentially facilitate the learning of either associations between intra-item features or association between intra-item features and the semantic identity of an object. That is, assessing the dissociability of top-down (i.e., the ability to retrieve intra-item features when given the identity of the object) and lateral associations required *prior learning* conditions that selectively boost the strength of each type of association. The flow of top-down information might be selectively enhanced by studying an object's name but not its visual details, whereas lateral associations might be selectively enhanced by studying an object's visual details without awareness of its identity. The key manipulation therefore relied on comparing participants' memory for objects that were studied as words as compared to words that were studied under continuous flash suppression (CFS: Tsuchiya & Koch, 2005). CFS entails simultaneously presenting a salient masking image (e.g., flashing, overlapping squares; referred to as a

'Mondrian'[3]) to a participant's dominant eye and a second image (e.g., a picture of an object) to their non-dominant eye. The masking image renders participants unaware of the image presented to the non-dominant eye. The CFS condition exposed participants to the visual details of the objects while limiting their awareness of object identity. The Word condition, in which subjects studied object *names* without CFS (i.e., participants were presented the word, no image, with full awareness), provided object identity but not the visual details of the tokens of objects participants would encounter at test. Additionally, a "No Study" baseline condition and a "Binocular Image" control condition, in which objects were studied as visual images without CFS (providing visual details as well as full awareness of object identity) were included.

To what extent does CFS prevent access to the semantic identity of objects? Previous studies have shown that CFS can provide an approximate way to prevent low-level visual information from being processed in a conjunctive manner. For example, participants are relatively insensitive to Gestalt information of items that are masked by CFS (Moors, Wagemans, van Ee, & De-Wit, 2016). Moreover, there is currently no strong evidence that the semantic information of images masked by CFS are retained by participants (Gayet, Van der Stigchel, & Paffen, 2014; Gelbard-Sagiv, Faivre, Mudrik, & Koch, 2016). Hence, it appears that CFS breaks the typically automatic association between the intra-item parts of an object and the identity of that object. Note that, for the CFS manipulation to be successful, it does not need to be the case that participants were completely unable to derive any semantic information from the objects that they

[3] The eponym Mondrian refers to the painter Piet Mondrian, given that the masking images resemble aspects of his paintings.

encounter under CFS. Even if it were the case that participants could derive non-specific information (e.g., whether an object was artificial or natural, whether the object is typically large or small), the CFS manipulation was only designed to limit participants' awareness of the identity of the objects that were encountered in this condition[4].

To complement these study conditions, two tasks were required that were differently sensitive to either the learning of intra-item associations or the learning of vertical associations. In a "Part-matching" task, participants indicated whether a pair of parts came from the same or different objects. In a "Naming" task participants attempted to provide the object's name from a visual part cue. In the Part-matching task, each part was a circular patch drawn from an object, chosen to be unrecognizable when viewed in isolation. Specifically, participants were given a two-alternative forced choice (2AFC) between a pair of patches (e.g., A_1 and A_2) that came from the same object, A, versus a pair of patches (e.g., A_1 and B_3) that came from two different objects, A and B. With a common patch in both choices (A_1), the task amounted to identifying which other patch (A_2 versus B_3) was associated with the common patch. In the Naming task, which was administered immediately after each 2AFC Part-matching test, participants were again presented with the common patch (e.g., A_1) and asked to name the object from which it was drawn. Analogously to how participants seem to know which dots in a Mooney image belong to the same object without knowing the identity of that object (van Tonder & Ejima, 2000), the 2AFC Part-matching task can be performed intuitively in the absence of object identification (e.g., perhaps familiarity enables them to determine that "those

[4] Although the processing that occurs for images masked by CFS during the masking may provide insight into definitions of different levels of consciousness (e.g., Dehaene & Changeux, 2011), that discussion is beyond the scope of this thesis. It is for this reason that we utilize the relatively neutral term "awareness."

patches just seem to go together"). In contrast, the Naming task requires access to object identity knowledge, but production of the identity can occur in the absence of lateral associations between separate parts of an object. To assess whether intra-item associations can be retrieved separately from top-down associations, we therefore tested for a dissociation between 2AFC Part-matching and Naming.

The dissociability of lateral associations from vertical associations was indexed by the following predictions. CFS was predicted to enable the learning of lateral, intra-item associations, but provide less learning (if any) of associations between these object parts and the objects' identities. Therefore, a dissociation between lateral and vertical associations required that study of objects masked by CFS boosted 2AFC Part-matching performance with little or no benefit to the Naming task. In the complementary Word study condition, participants studied with full awareness only the word-name of each object. This condition was predicted to support the learning of object identities, which may influence subsequent test performance via top-down associations (i.e., from object identity to part A_2) but was not predicted to support learning of bottom-up or lateral visual associations (because there is no viewing of the subsequently tested object parts). Therefore, a dissociation between lateral and vertical associations also required that study of the object boosted Naming performance with little or no boost to performance in the 2AFC Part-matching task. These boosts are relative to the No Study condition, and performance in the Binocular condition was predicted to be highest in both tasks.

2.2 Methods

2.2.1 Participants

A minimum of 1000 data points per condition were sought. Participants were recruited through the University of Massachusetts, Amherst's SONA account and campus fliers. In total, 53 participants completed the experiment. Participants were compensated either with course credit or at a rate of 10$ per hour.

2.2.2 Materials

Images of 144 unique, everyday objects (e.g., chair, couch, monitor) were gathered from the internet. Images of objects were presented in gray-scale, cropped and resized (maintaining the aspect ratio of the original object) such that the longest cardinal axis (horizontal or vertical) of the object extended to 295 pixels. Objects were superimposed upon a gray (RGB values: [161, 161, 161]) background of 300x300 pixels and framed by a 150 pixel white border. They were displayed on a 24" LED monitor at a resolution of 1920 x 1080 (120 *Hz* refresh rate).

Object patches were created using three aperture views (circles of radius 80 pixels) of each object. The view through an aperture could potentially include some of the background as well as part of the object. The choice of aperture location was constrained such that the ratio of object to background visible through each aperture was no less than the ratio of the object to the background in the entire 300x300 pixel image. Fixed pairings of the objects were determined in advance (creating 77 pairs from the 144 objects), with each pairing chosen to yield a pair of mismatched apertures (i.e., apertures from two different objects to be used as the foil stimulus in the 2AFC task) that were not

obviously mismatched. Each pair of objects was used to create two different intact-rearranged 2AFC trials, with both trials created from a given object pair being assigned to the same study condition. For instance, object A had three apertures (A_1, A_2, and A_3) and object B had three apertures (B_1, B_2, and B_3). From these apertures, two different 2AFC Part-matching trials were created: A_1-A_2 (intact) versus A_1-B_3 (rearranged), and B_1-B_2 versus B_1-A_3. Both trials were then assigned to the same condition (e.g., "Binocular").

Mondrian masks for CFS were constructed from overlapping, colored rectangles. The rectangles were presented in a cropped, 300x300 pixel square, centered on the presentation screen. The Hue, Saturation, and Value (HSV space) of the colors of the squares were determined by uniformly random samples between [1/6, 1]. The width and height of the rectangles were randomly chosen to be between 15-40 pixels. 1000 such rectangles were drawn for each mask.

2.2.3 Procedure

Experimental code was written with MATLAB (MathWorks, 2015) scripts using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007). Prior to the main experiment, ocular dominance of each participant was measured via the Porta test (as performed by Roth, 2002). Figure 3 presents an overview of the main experimental procedure.

The experiment contained nine blocks, each with a study phase and test phase. Participants encountered 16 objects per block (8 pairs of objects), with each study trial presenting just one of the 16 objects. The two objects in each pair were assigned to the same study condition. The assignment of object pairs to blocks was fixed across participants, but the study condition for each object pair was randomly determined for

each participant, with the constraint that each block contain 4 test trials for each of the 4 conditions. The study conditions were (1) No prior study, (2) Word (the name of the object was presented in text form, rather than the image), (3) CFS (the image of the object was presented, masked by CFS), and (4) Binocular (the image of the object was presented without CFS masking).

2.2.3.1 Study Phase

During the study phase of each block, participants saw 12 objects, one at a time (4 from each of the 3 conditions that contained prior study). The order of the 12 presentations was pseudo random, disallowing more than 3 presentations of the same study condition in a row. Participants studied the 12 items in the same order twice, with the second presentation of the list of items immediately following the first. Each study trial began with a $500\ ms$ fixation square, followed by a CFS mask (presented binocularly), with a $150\ pixel$ white border. On Word trials, the name of the object gradually appeared (linear ramping of transparency across $1000\ ms$) in front of the mask, presented to both eyes. On Binocular trials, the image of the object gradually appeared in front of the mask, presented to both eyes. On CFS trials, the image of the object gradually appeared in front of the mask for the participant's non-dominant eye, whereas the dominant eye continued to view only masks.

Presentation of different images to the different eyes was accomplished by using NVidia 3D "shutter" glasses, which synchronize with the display monitor to allow only one eye to see through the glasses on a given refresh cycle; by alternating between two displays on each refresh cycle, each eye views a different display. On all trials, once the

object or word reached full opacity it was presented for $2500\ ms$, and then gradually decreased to full transparency over $500\ ms$.

Throughout study trials, the Mondrian masks changed every $100\ ms$ (i.e., at $10\ Hz$), by altering the position, size, and color of the rectangles in the mask. A set of masks was randomly created for each participant, and this same set was used for all trials for that participant, with each trial presenting the masks in the same order. Each $100\ ms$ duration mask was presented only once per trial.

At the start of each CFS study trial, participants were instructed to press a button as soon as they detected that an object was present. They were instructed to give this response only if they were confident that there was an object, but before they could identify the object. As soon as they gave this response, the trial ceased so as to minimize learning with awareness. To encourage fast responses, they were shown their reaction time. To encourage accuracy, $1/3$ of study trials were catch trials, in which no object appeared. If they responded on catch trials, participants saw the message: "CAREFUL! No object appeared". This self-termination of study on CFS trials was included to: 1) provide the maximum opportunity for learning with limited awareness; and 2) control the degree of awareness. Regarding the second rationale, a fixed duration of study would likely result in a mixture of trials, with some experiencing full "breakthrough" from CFS and thus full awareness of object identity, while other trials would be completely without any awareness that anything was shown. By allowing participants to self-terminate study at the first stage of breakthrough (i.e., when they first started to become aware that an object was present), the aim was to place all study trials at the same level of limited awareness.

At the start of Word trials, participants were instructed to “please imagine the following object in detail, as if it were presented over the flashing squares.” At the start of binocular trials, participants were instructed “please study the details of the following object.”

At the end of every study trial for all study conditions, participants were asked two questions about the object. First, they gave a four-valued Perceptual Awareness Scale (PAS) (Ramsøy & Overgaard, 2004) rating, with values 0-3 (these data were collected but analyses are not presented). Participants were instructed to use the scale as follows: “If you CLEARLY SAW something besides the squares, AND COULD NAME IT, answer 3”, “If you DEFINITELY saw something, but are unsure what (though you might be able to guess), answer 2”, “If you only POSSIBLY saw something, but COULDN'T accurately say what it was, answer 1”, and “If you didn’t see anything besides the squares, answer 0.” These instructions were presented at the beginning of the experiment, and a reminder was presented on all study trials. The second question was included to encourage attention to the visual details of each object (in the word condition, the imagined visual details of the object). The question was one of four randomly assigned questions: 1) “Was the object symmetric across its horizontal axis?”, 2) “Did the object fill more than $1/4$ of the flashing squares?”, 3) “Was the object reflective?”, or 4) “Did the object contain multiple parts?”

2.2.3.2 Test Phase

Each test phase occurred immediately after the corresponding study phase. The test phase was self-paced, and the 16 items were presented in random order. On each test trial, participants were first asked to choose which of two pairs of parts came from the

same object. Immediately after each 2AFC Part-matching test, the Naming test for the corresponding object occurred (i.e., the two test formats were interleaved on a trial-by-trial basis). Specifically, the part that was common to both forced-choice options reappeared by itself, and the participant was asked to name the corresponding object. In both tasks, participants were encouraged to "use your memory from the items that you studied, if that helps."

For all participants, the first block was used as a practice. The practice block was excluded from all analyses. The entire experiment lasted approximately 2 hours.

2.2.4 Analyses

State-trace analyses were the primary method to infer whether there was a dissociation between lateral and vertical associations (Bamber, 1979; Dunn, 2008; Kalish, Dunn, Burdakov, & Sysoev, 2016). State-trace analyses are motivated in the following section and an overview of the mechanism is provided. Further details of state-trace analyses are presented in Part 2 of this thesis, as well as a novel technique for performing a state-trace analysis.

2.2.4.1 Overview of State-Trace Analysis

The predicted pattern of results outlined above would constitute a double dissociation (i.e., one manipulation selectively affects one performance measure, while a different manipulation selectively affects a different performance measure). Such a finding could provide evidence that lateral, intra-object associations can be learned separately from top-down influence of object identity information. However, this evidence would not be clear-cut – double dissociations can arise as an artifact of

comparing two performance measures that differ in sensitivity to a single underlying representation (Bamber, 1979; Dunn & Kirsner, 1988; Loftus, 1978; Newell & Dunn, 2008; Wagenmakers, Krypotos, Criss, & Iverson, 2012). That is, a double dissociation does not decisively exclude the possibility that performance on both the 2AFC Part-matching and Naming tasks relies on a single type of memory representation (e.g., the associations between visual parts and object identity).

To address this concern, a state-trace analysis was performed (e.g., Bamber, 1979; Kalish et al., 2016). A state-trace analysis directly tests whether more than one latent variable contributes to a given behavior, identifying dissociations in a manner that is not rendered ambiguous by possible differences in task sensitivity. A state-trace analysis frames the test in terms of an independent factor (e.g., 4 study conditions), two dependent variables (e.g., implicit 2AFC versus explicit naming), and either one (e.g., memory strength) or two (e.g., separate explicit and implicit object knowledge) latent psychological variables of interest. The *outcome space* (Dunn, 2008, pg 429) is defined as all possible combinations of values for the two dependent variables (in the current case, the values are probabilities, so the outcome space is a unit square). The *state-trace* (Dunn, 2008, pg 429) describes the subset of the outcome space that is feasibly reached by a model that maps the independent factor via latent variables, into the outcome space.

The state-trace analysis asks whether the ordering of the average performance across conditions is the same on each of the two measures. That is, can the observed data be described by a single monotonic function, or does the ordering of performance on the two conditions differ between measures? If the data can be fit with an arbitrary monotonic function, then the data may have been produced by either a model with one or

more latent variable (Ashby, 2014; Dunn, Kalish, & Newell, 2014). In that case, the data do not provide explicit evidence for either kind of model, but parsimony could be invoked to favor the simpler model. However, if the data cannot be fit with a monotonic function, the unidimensional model may be rejected, lending credibility to the bidimensional model.

A state-trace analysis technique developed by Kalish et al. (2016) was used to test for a dissociation. Their procedure first finds the best-fitting, single-latent-variable model. This model is defined such that performance on the two tasks (i.e., 2AFC Part-matching and Naming) depends on a single kind of representation (e.g., "memory strength"). This definition is instantiated by a requirement that the rank ordering of the estimated performance in all experimental conditions should be the same in both dependent measures. This instantiation of the effect of a single latent variable (i.e., identical rank orderings of performance in both conditions) follows from an additional assumption that the latent variable has a monotonic effect on performance. For example, assume that performance on the two tasks depended on a single latent "memory strength," and that there were three conditions, 1, 2, and 3, that resulted in low, medium, and high memory strength, respectively. Since performance depends only on the value of the memory strength, it could not be the case that, for the first measure, performance in condition $1 < 2 < 3$ but for the second measure condition $1 < 3 < 2$.

Next, the data are fit with a model that allows for different rank orders for each dependent measure. This corresponds to the assumption of two or more latent variables (e.g., two separate representations on which memory performance relies, such as lateral, intra-item associations separate from top-down associations). Finally, the fits of these two

models are compared using bootstrap resampling (Wagenmakers, Ratcliff, Gomez, & Iverson, 2004), which penalizes a model for being too flexible and thus fitting noise (in this case, the single-latent-variable model is less flexible). If the single-variable model results in a significant loss of fit, then a single-representation account of the data can be rejected (Kalish et al., 2016), demonstrating a dissociation without concern over possible differences in sensitivity of the two dependent measures.

Kalish et al. (2016) outlined statistical tests that can be performed on either the participant-level average performance in each condition or a binomial version that can be used to assess monotonicity within individual observers. Both tests exhibit tradeoffs. We opted to analyze each observer's average performance, which assumes only that observers' average performance within each condition on each task is normally distributed.

2.2.4.2 Correction for Guessing

One potential concern with this experimental procedure is that performance for objects that were studied with awareness (i.e., the word and binocular conditions) may have artificially boosted naming performance as a result of a guessing strategy based on explicit memory for the list of previously studied objects. That is, observers could adopt a strategy whereby whenever they encountered an aperture that they did not recognize, they simply guess from among the object names that they remembered studying.

To correct for this potential strategy, it was assumed that correct performance on a given trial in the Naming task could arise either because the observer correctly identified the whole object from the part using their memory (either from prior study or pre-experimental perceptual knowledge), or correctly guessed the name. That is, the

probability of naming an object on a given trial, $P(name)$, was assumed to follow the equation

$$P(name) = P(name_{memory}) + \left(1 - P(name_{memory})\right) * \left({P(name_{guess})}/{4}\right)$$

where $P(name_{memory})$ is the probability of naming the item using memory and $P(name_{guess})$ is the probability of guessing an item from the most recent study list that happened to come from the same study category (i.e. the same experimental condition) as the correct name. Thus, in this equation, the $P(name_{guess})$ does not reflect overall guessing rate but is instead the condition-specific guessing rate[5]. With this study condition-specific definition, $P(name_{guess})$ is divided by 4 because for each test item, there were exactly 4 items of that condition in the most recent study list (i.e., if memory failed, participants had a one in four chance of guessing correctly if they only guessed objects that were studied in the same condition as object they were trying to guess). $P(name)$ is the observed proportion of objects that were correctly named, but $P(name_{memory})$ is the value of interest. $P(name_{guess})$ must had to be estimated.

To estimate $P(name_{guess})$ for each observer in each condition, we first tallied the number of trials in which the observer provided neither the correct name, nor the name of the partner object in the pair. Trials were excluded in which the partner object was named because this name may have been prompted by the appearance of the partner

[5] The condition-specific guessing rate was used instead of the overall guessing rate because this supplementary analysis was only concerned about the extent to which guessing artificially increased naming performance in some conditions but not others. That is, the concern is not that guessing occurred, but that guessing may have helped more in some conditions as compared to others. For this reason, the guess rate needed to be calculated in a condition-specific manner.

object aperture in the 2AFC task that was immediately prior to the naming task (and so including these trials would have biased the estimate of $P(name_{guess})$). Of these trials, we determined the condition of the name that was provided. That is, we tallied the number of times the produced incorrect name was the name of an object that was studied as a Word on the most recent list, studied as a picture under CFS on the most recent list, studied as a picture binocularly on the most recent list, one of the objects chosen for the not studied condition on the most recent list, or was not on the list at all (e.g., from a different list, or a random guess). Each of these values was divided by the total number of error trials in the condition of interest (trials in which neither the correct name nor the partner name were produced). Substitutions were categorized as coming from one of the study conditions only if the name provided was from an object encountered during the current block of 16 objects.

The resulting four rates (including only the substitutions with objects on the current study list) provided an estimate of $P(name_{guess})$ for each condition. These values were entered into the above equation, to derive values for $P(name_{memory})$. Correcting for guessing did not change the conclusions of the state-trace analysis.

2.2.5 Results

To assess the effect of study condition on 2AFC Part-matching and Naming performance, separate (i.e., one for 2AFC and one for Naming) linear mixed models were fit using the lme4 (Bates, Mächler, Bolker, & Walker, 2015) R (R Development Core Team, 2016) package, treating condition as a fixed-effect and giving each participant a random intercept (Figure 4). Post-hoc comparisons revealed that all pairwise comparisons

were significantly different (all $p < 0.05$, Tukey's method for correcting multiple comparisons), but that the Word and CFS conditions were reversed for the 2AFC Part-matching task relative to the Naming task. The average 2AFC Part-matching accuracy for an object studied as a Word was 0.73 (SE = 0.014), which was lower than the average 2AFC accuracy for an object studied under CFS (M = 0.77, SE = 0.014; $t(6728) = 2.868$, $p = 0.022$). In contrast, objects studied as words were more often named (M = 0.42, SE = 0.02) as compared to objects studied under CFS (M = 0.37, SE = 0.02; $t(6728) = 3.221$, $p = 0.007$).

This dissociation between the effects of studying the identity of an object and studying the visual details of an object suggests that the study conditions may have allowed learning of two distinct kinds of information. Furthermore, these two distinct kinds of information cannot be top-down (e.g., learned during Word study) versus bottom-up (learned from visual study), because a benefit on the Part-matching task from enhanced bottom-up processing, in the absence of lateral associations, can only be revealed via the "top" (i.e., object identification, in which case Naming performance would have been equally enhanced).

However, as mentioned above, these statistics are inherently ambiguous with respect to inferring how many cognitive variables are required to account for the data (Bamber, 1979; Dunn & Kirsner, 1988; Newell & Dunn, 2008). For example, 50% accuracy in the two tasks may imply different levels of performance (i.e., given that chance in the 2AFC was 50%, but chance in the Naming task was presumably much lower), or than an increase in 10% performance as compared to stimuli that were not studied in the two tasks may imply different levels of learning. Therefore, a state-trace

analysis was conducted to ask whether lateral, part-to-part associations were dissociable from top-down associations between object identity and visual parts.

The technique developed by Kalish et al. (2016) was used to conduct a state-trace analysis. In these data, the single-latent-variable model fit significantly worse ($p = 0.0027$; Figure 5), indicating that performance on the 2AFC Part-matching and Naming tasks depended on more than one latent variable.

2.2.6 Discussion

This first experiment was designed to probe whether the learning of lateral associations between visual parts of objects could be dissociated from the learning of vertical associations. A state-trace analysis was used to determine whether performance on the 2AFC Part-matching and Naming tasks could have depended on just a single latent representation. The key result was whether performance could be accounted for by a unidimensional model in a state-trace analysis. The unidimensional model was rejected by the state-trace analysis technique developed by Kalish et al. (2016). This rejection suggests that performance in the two tasks depended on more than a single latent variable. Given the setup of the experiment, it is inferred that CFS provides a way to preferentially boost the learning of intra-item associations as compared to part-whole associations.

However, the state-trace analysis technique developed by Kalish et al. (2016) does rely on a – possibly incorrect – assumption: that there is no trial-level correlation between performance on the two tasks. Part 2 of this experiment will detail how correlation between the two dependent measures could lead to an incorrect conclusion

about the latent dimensionality of the cognitive system in question and provide a solution based on a Bayesian mixture model of bivariate probits.

CHAPTER 3

HIERARCHICAL BAYESIAN STATE-TRACE ANALYSIS

3.1 Introduction

The outline of this part is as follows. First, a reminder of key terminology of State-Trace analyses will be presented, more detailed description of the theory of state-trace analysis will be presented. The method proposed in this thesis will follow. To validate the technique, data were simulated from monotonic and non-monotonic models and the technique was used to recover the data-generating model. The penultimate section will apply the novel technique to the data collected in Part 1. A final section will discuss a few remaining limitations to the technique.

3.2 Detail of State-Trace Analysis

State-trace analyses provide a way to test the dimensionality of psychological spaces (Bamber, 1979; Dunn, 2008). A state-trace analysis frames a test in terms of an independent factor (i.e., conditions), two dependent variables (e.g., accuracy in cued-recall or recognition), and either one or two latent psychological variables of interest (e.g., memory strength or attentional weight). The *outcome space* is defined as all possible combinations that values of two dependent variables could take. The *state-trace* describes the subset of the outcome space that is feasibly reached by a model that maps an independent factor (i.e., experimental conditions) via latent variables, into the outcome space. The goal of a state-trace analysis is to determine which model's state-trace best accounts for the observed state-trace – where 'best' is a measure of a model's ability to recapitulate the data at hand, penalized by its flexibility. This is analogous to most

experiments, but the insight of a state-trace analysis is to frame the models in question in a general enough way so that only their latent dimensionalities differ. That is, a state-trace analysis provides a way to make inferences on the dimensionality of the latent space without precise specification of the transformation from latent to outcome space (Figure 6). Although the state-trace analysis can be performed on models with an arbitrarily sized latent space, typically the models under comparison have either just one or two latent dimensions. Comparisons between such *unidimensional* and *bidimensional* models will be the focus of this paper.

Although models are only defined in terms of the dimensionality of the latent variables, this relatively general definition enables the models in question to make different, testable predictions. Specifically, the unidimensional model predicts that the data as plotted in the outcome space will exhibit a monotonic relationship[6]. Hence, if the data resemble a monotonic function, then the data may have been produced by either a unidimensional or bidimensional model (Ashby, 2014; Dunn et al., 2014). However, if the data cannot be fit with a monotonic function, the unidimensional model may be rejected, lending credibility to a bidimensional model. In such a scenario, it is then up to the researcher (analogously to the specification of an alternative hypothesis in null-hypothesis significance testing) to specify which bidimensional provides the most plausible account of the data.

[6] Note that this prediction only holds because of the relatively mild assumption that f and g (in both models) are themselves monotonic. This translates to an assumption that, for example, if an increase a latent variable like memory strength from low to medium elicits higher recognition accuracy, an increase in memory strength from medium to high will not correspond to a *decrease* in performance on recognition.

In the following explanation, an observed instance of the dependent variables will be referred to as x and y, respectively. Similarly, a value of the latent parameters A and B will be referred to as a or b, respectively. The transformation of these latent variables (the model) into outcomes will be written as functions, f and g. In a unidimensional model, outcomes are separate monotonic functions of a single latent variable

$$x_u = f_u(a),$$

$$y_u = g_u(b).$$

A hypothetical state-trace of a unidimensional model is plotted in Figure 6a.

In a bidimensional model (Figure 6b), the outcomes depend on two latent variables. The resulting state-trace is defined by

$$x_b = f_b(a, b),$$

$$y_b = g_b(a, b).$$

One impediment to conducting State-Trace Analyses has been the lack of a general solution for the practicalities of selecting between competing models; although idealized models make different predictions for performance, the statistical tests that should be used in the face of measurement noise remains an open question. That is, although state-trace analyses provide a way to distinguish between models, the results of an experiment are inevitably better described by

$$\boldsymbol{x} \sim \boldsymbol{f_u}(\boldsymbol{\theta_u^1}) \quad \boldsymbol{x} \sim \boldsymbol{f_b}(\boldsymbol{\theta_b^1})$$
$$\boldsymbol{y} \sim \boldsymbol{g_u}(\boldsymbol{\theta_u^2}) \quad \boldsymbol{y} \sim \boldsymbol{g_b}(\boldsymbol{\theta_b^2})$$

where f, g are unspecified distributions with parameter vectors θ (indexed by response, some (but not necessarily all) of whose elements will be functions of either a, b, or both), and the symbol, $\sim$, reads "is distributed as." The challenge is in how to determine the better fitting model when f and g are an unknown distribution.

Methods for handling the noise inherent to experimentation have been proposed (Bamber, 1979; Davis-Stober, Morey, Gretton, & Heathcote, 2016; Dunn & James, 2003; Dunn & Kirsner, 1988; Kalish et al., 2016; Loftus, Oberg, & Dillon, 2004; Pratte & Rouder, 2012; Prince, Brown, & Heathcote, 2012). Although these methods provide tools to conduct a State-Trace Analysis, none of these methods provides a way to estimate correlations of the condition effects between performance on the dependent measures. These methods are based on the differences in expected performance in the two tasks, but unmodeled correlations between the two measures could skew these differences (Figure 7). For example, on any given trial, the observer may decide to devote more effort to one of the two tasks, with the choice of which task to favor varying from trial to trial, leading to a negative dependency between the two measures. If a trial-level negative dependency between task outcomes existed, the distributions of performance on the two conditions may appear to lie on two separate state-traces. However, the underlying truth could be that a single-state trace simply exhibits this negative correlation, and so the unidimensional model is relatively handicapped. The problem in question, then, is how to adjudicate between the following models

$$(x, y) \sim h_u(\theta^1)$$
$$(x, y) \sim h_b(\theta^2)$$

which include the bivariate distributions h_u (for the univariate model) and h_b (for the bivariate or multivariate model), not necessarily of the same form, that may have different sets of parameters (θ^1 vs. θ^2) and whose dimensions may be correlated.

3.3 Hierarchical Multivariate Probit Model for State-Trace Analysis

A graphical description of the style utilized by Kruschke (2014) is presented in Figure 8. The 2AFC and Naming tasks were modeled with a hierarchical Bayesian, bivariate probit model (Greene, 2017; Stan Development Team, 2017b). The model is bivariate to reflect the two test questions (i.e., 2AFC and Naming). This bivariate probit can be seen as a bivariate extension of an Equal Variance Signal Detection model in which the criterions for both decisions have been fixed to 0 (Macmillan & Creelman, 2005). However, no specific cognitive interpretation is meant to be implied by the dimensions of this regression model; the two dimensions should not be interpreted as two separate kinds of, e.g., familiarity. Put another way, the model was not designed to estimate the underlying latent variables that gave rise to behavior; both the unidimensional model and bidimensional model rely on the bivariate probit because both models need to account for performance in two tasks. Instead, the two dimensions of this model can be thought of with the relatively atheoritic terms "evidence required to make a correct response."

However, intuitions for the mechanisms of a signal detection model are useful for understanding the behavior of this bivariate probit model. The model assumes that performance on the two tasks is a thresholded process whereby a participant will get a correct response when the value of evidence, z, is above some threshold (fixed to 0). Each pair of observations corresponds to a single trial, defined by a unique combination of a participant, s encountering an item, i, that was studied in condition, c. For example, both z above 0 corresponds to a trial in which the participant gets both tasks correct.

The means of the data-generating distribution were the sum of item-, participant-, and condition-effects. Intuitively, each of these effects can be thought of as their respective propensity to elicit the correct response for each trial. For example, an item-effect corresponds to the guessability of an item encountered in the 2AFC when it was not studied condition and the nameability of that unstudied item in the Naming task (given that the model is set up to treat the not-studied condition as a reference group).

Each item-, participant-, and condition-effect was sampled from bivariate normal distributions. These bivariate normal distributions were each parameterized by a vector of means describing the magnitude of the effects (μ), a correlation matrix describing the correlation between the two effects (Ω), and standard deviation along each dimension (τ). The means were fixed to 0. The item- and participant-effects were treated as random, and the condition-effects were treated as fixed (Vandekerckhove, Tuerlinckx, & Lee, 2011). That is, all item- and participant-effects were sampled from two bivariate normal distributions, representing the population distributions for item- and participant-effects, respectively. Each of the four condition-effects were sampled from a separate bivariate normal, allowing for trial-level positive or negative dependencies between the two questions, with this dependence potentially differing between conditions. Non-centered parameterizations of the bivariate normal for the item- and subject-effects were used (Betancourt & Girolami, 2015).

Distributions at the top of the figure constitute the subjective hyperpriors for the model. The correlation matrices were all given LKJ hyperpriors (Lewandowski, Kurowicka, & Joe, 2009) ($LKJ(1.5)$), and the variance parameters were given gamma

hyperpriors ($\Gamma(2,4)$). In total, the hyperpriors included three LKJ and three gamma distributions.

To explain how to incorporate the monotonicity and non-monotonicity constraints, it is useful to describe the assumed data-generating distribution as an equation. The model specifies that outcome, y, on a given trial is modelled as a threshold process

$$\binom{x}{y}_{c,s,i} = \Phi^2\left(\binom{\mu_x}{\mu_y}_{c,s,i}, \Omega_c\right)$$

.

The function Φ^2 is the cumulative distribution function of the standard bivariate normal distribution (calculated as described by Greene, 2017), which serves to take the unconstrained parameters μ into the range of 0-1 (corresponding to the likelihood of either an incorrect or correct response). Although the standard deviations of the bivariate normal implied by the function Φ^2 are restricted to be one[7], correlation between the two dimensions is modelled through the matrix Ω_c. A main parameter of interest is the effect of each of the study conditions, which appears in the following equation

$$\mu_{c,i,s} = subject\_intercept_s + item\_intercept_i + \beta_c$$

.

which says that each trial is the summation of subject s's ability, item i's guessability, and the effect of condition c (which is given the label β). That equation would be repeated once to give both μ_x and μ_y.

[7] Much like the common practice in a Signal Detection framework to specify that the distribution of familiarity strengths has a standard deviation of 1, working with the cumulative density function of only the standard bivariate normal does not sacrifice generality.

The equations so far emphasize the data generating process, describing how the model describes what happens on a given trial. To model the monotonic vs. non-monotonic models, the following equation is useful, which implies the same model but instead emphasizes the probability mass associated with a given observation

$$p\left(\binom{x}{y}_{c,s,i} \mid \binom{\mu_x}{\mu_y}_{c,s,i}, \Omega_c\right) = biprobit\_pmf\left(\binom{x}{y}_{c,s,i} \mid \binom{\mu_x}{\mu_y}_{c,s,i}, \Omega_c\right)$$

which shows that the model determines the likelihood of a trial based on a bivariate probit distribution parameterized by a pair of means and a correlation matrix. That is, the likelihood of an observation given the parameters of the model, $p\left(\binom{x}{y}_{c,s,i} \mid \binom{\mu_x}{\mu_y}_{c,s,i}, \Omega_c\right)$, is a function of the bivariate probit probability mass function, which is equivalent to Φ^2. Pairs of correct or incorrect responses (observations) are modelled jointly by the bivariate probit.

In the model described so far, the condition-, subject-, and item-effects are all unconstrained. This model is more flexible than either the monotonic and non-monotonic models. To constrain the model requires imposing order constraints on the condition effect, β. The different types of order constraints will imply either the monotonic or non-monotonic models. For example, one monotonic model can be realized by requiring that the condition parameters are in the following order: Not Studied < Word < CFS < Binocular, for both the effect of 2AFC and Naming performance. A non-monotonic model could be instantiated by enforcing the condition effect to have that order for the 2AFC task, but then enforcing the order of Not Studied < CFS < Naming < Binocular in the Naming task. The dual challenges of both enforcing monotonicity and specifying all monotonic or nonmonotonic orderings in single models is addressed next.

Monotonicity is enforced using the method presented by Bürkner (2017). In each dimension of β, a single parameter β^{raw} models the direction and magnitude of the direction and magnitude of the experimental manipulation. A separate parameter, ζ, models the normalized distance between conditions

$$\beta_c = \beta^{raw} \sum_{i=1}^{n\_conditions} \zeta_i$$

Note that there is a ζ for each condition, and the ζ for each condition is a vector. Additionally, ζ is required to be a simplex (i.e., requiring that $\zeta_i \in [0,1]$ for all i and $\sum_{i=1}^{n\_conditions} \zeta_i = 1$). Any distribution capable of generating a simplex could be a suitable prior for ζ. In this case, a two-stage method is used[8]. First, $n\_conditions - 1$ intermediate parameters, ζ^{raw}, are sampled from a normal distribution

$$\zeta^{raw} \sim N(0, \sigma_\zeta),$$

once for each ζ in each condition. The standard deviation of that normal distribution, σ_ζ, controls the extent to which the prior prefers an equal mixture of orderings as compared to preferring one single ordering[9]. Then, the ζ^{raw} are concatenated with a single 0, and the resulting vector (which the following equation still refers to as ζ^{raw}) is transformed with the softmax function in the ζ for a given condition

$$\zeta_c = \frac{\exp(\zeta_c^{raw})}{\sum_{i=1}^{n\_conditions} \exp(\zeta_i^{raw})}$$

[8] The application of a softmax to intermediate parameters sampled from a normal distribution was chosen rather than sampling from a Dirichlet for computational reasons. Initial attempts to sample θ directly from a Dirichlet often resulted in highly autocorrelated samples of the λ parameter, and informal comparison between use of the Dirichlet and the two-stage procedure suggested that the two-stage method resulting in more efficient sampling.

9 In all uses of the model in this paper, the σ_ζ was set to 2.

The softmax transforms the unbounded ζ^{raw} into a simplex. Note that one fewer ζ^{raw} are sampled as there are allowable orders in a model because the softmax function is invariant to equal scaling of the inputs (e.g., softmax of the vector [1,1,1] is equal to the softmax of the vector [2,2,2]). Without pinning one of the values of ζ^{raw}, the ζ parameter would be non-identifiable (e.g., pinning the first element of the previous example gives the vectors [0,1,1] and [0,2,2], which when passed through the softmax function produce different vectors).

To define a single, full monotonic model (which includes all sets of monotonic orderings) and a single nonmonotonic model (which encompasses the monotonic orderings), we follow (Davis-Stober et al., 2016) in using a mixture mixture-model approach. The number of components in the mixture is therefore the $m * k!$ possible permutations, where $k = 4$ is the number of conditions and $m = 1$ in the monotonic case and $m = 2$ in the non-monotonic case.

The parameter λ_j is now introduced, whose interpretation is the probability that the data were generated by condition effects with the order, j. Note that λ can be interpreted as the probability of a given ordering being true, in that it is also a simple ($\sum_{j=1}^{n\_orders} \lambda_j = 1$). The likelihood of a given trial, incorporating all of the n_orders that are allowed by a given model is then

$$P\left(\binom{x}{y}_{c,s,i} | \binom{\mu_x}{\mu_y}_{c,s,i}, \Omega_c\right) = \sum_{j=1}^{n\_orders} \lambda_j \, biprobit\_pmf\left(\binom{x}{y}_{c,s,i} | \binom{\mu_x}{\mu_y}_{j,c,s,i}, \Omega_c\right)$$

.

The preceding equation implies that the likelihood of the observed data is a mixture (as weighted average) of as many bivariate probit probability mass functions as there are orders under consideration. The components of the mixture are differentiated by how their μ parameters were constructed (the order constraints put on the condition effects). The weights associated with each ordering are now parameters to be estimated. Put another way λ instantiates the uncertainty associated with the true order of condition effects, and the variety of orders that are allowed by a given model. The same two-stage approach that models the ζ was used to model λ.

Inference can then proceed through model comparison. Both the monotonic and non-monotonic model are fit once (where the two models are differentiated based on the orders of condition effects under consideration). Then, any suitable criterion for picking the best fitting model can be used. In the present paper, Watanabe-Akaike Information Criterion (WAIC) is chosen (Vehtari, Gelman, & Gabry, 2017; Watanabe, 2010; Appendix B).

For computational reasons, one final adjustment was used in the simulations and application below. That is, infeasible orderings were not included. In the particular case of Experiment 1, it would be highly improbable for items encountered in the Not Studied condition to be either named or chosen in the 2AFC more often than any of the other conditions. Likewise, items that were studied Binocularly are surely the most likely to be named and chosen correctly. As such, rather than considering, for example, $4! = 24$ possible orderings in the monotonic model, the only orderings that are considered are those in which β for the Not Studied condition is lower than for the other three and β for the Binocular condition is higher than for the other three. This reduces the number of

orderings in the monotonic model from 24 to 2, and in the non-monotonic model from 48 to 4. This reduction in the number of orderings considered may also be justified from a Bayesian perspective; the models considered express the strong prior belief that studying an object in any of the three study conditions will never decrease task performance, and that studying an image of object will full awareness enables more learning than the other two study conditions.

All models were fit using the No-U-Turn Sampling algorithm as implemented by the R interface to the Stan language (Stan Development Team, 2016, 2017a).

3.4 Model Recovery Via Simulation

To ensure the robustness of the proposed technique, it is beneficial to know that the technique can recover the correct model in simulated data. That is, when the data are generated from a monotonic model, the technique should pick the monotonic model. When the data are generated from a non-monotonic model, the technique should pick the non-monotonic model. This section provides a minimal validation of the technique, in particular demonstrating the technique's robustness to correlation between the two measures.

To assess the robustness of the proposed technique to correlation between the two measures, three pairs (one monotonic and one nonmonotonic) of datasets were simulated with varied levels of correlation (Figure 9). Each pair consisted of a simulation from the model described in the previous section, once for an ordering that could be explained by a unidimensional model (which could therefore have also been produced by a bi-dimensional model) and once for an ordering that could be explained by a bi-dimensional model (which could not have been produced by a unidimensional model). To probe how

the technique performs with data that are minimally, moderately, and highly correlated the levels of correlations considered were 0, 0.5, and -0.9.

Each dataset included 50 participants presented with 40 items in each of 4 conditions (i.e., 8000 observations in total). The remaining parameters for the simulated datasets were specified in the unconstrained space (i.e., as probit values). The condition effects in the monotonic model were $\big((-1,-1),(-0.5,-0.5),(0.5,0.5),(1,1)\big)$, and for the non-monotonic model they were $\big((-1,-1),(-0.5,0.5),(0.5,-0.5),(1,1)\big)$. The variability of participant and item effects was set to 0.5 probits (i.e., participant and item effects were sampled from a bivariate normal distribution with no correlation and standard deviation of 0.5 along both dimensions). These parameters produced data from each condition that was highly overlapping (Figure 9).

In all six comparisons, the correct model was chosen by more than 3 standard errors (Appendix B). This is encouraging, as it suggests that the technique can pick the correct model given sufficient power.

To explore the limits of the technique, additional smaller datasets were simulated that included that had 1) only three conditions, 2) only 20 items per condition, 3) either 10, 20, or 30 participants, 4) correlations of -0.5, 0, or 0.5, 5) nine different arrangements of condition effects spanning from monotonic to minimally nonmonotonic, and 6) the nonmonotonic model was divided into two model – a "full" model that contained both monotonic and nonmonotonic orderings and a true nonmonotonic model that did not contain monotonic orderings (the monotonic model was also fit to these datasets, resulting in three fits per dataset). However, the WAIC was equivocal in most of these comparisons. That is, the difference in the WAIC for most of these comparisons was less

than 2 standard errors. This is, of course, the desired response when the data were monotonic. However, it does suggest that the technique may require many observations to have the power to detect to nonmonotonicity, depending on the degree of monotonicity and the variability due to items or participants.

3.5 State-Trace Analysis Applied to Experiment 1

Applying this novel technique to the results from experiment 1, 6 chains were run, with 1000 samples of warmup and an additional 1000 samples from the posterior (6000 posterior samples, overall). The $split - \hat{R}$ for all parameters in both models was below 1.1 and there were no divergent transitions (Appendix A; Betancourt, 2017; Betancourt & Girolami, 2015; Gelman, 2014; Gelman & Rubin, 1992).

The monotonic and non-monotonic models were fit to the data and model fits were compared via WAIC (Vehtari et al., 2017; Watanabe, 2010). WAIC was calculated using the log-likelihood of all 6000 samples from the posterior samples. The non-monotonic model ($elpd_{WAIC}$[10] -7882.3, se 44.4) was preferred to the monotonic model ($epld_{WAIC}$ -7890.4, se 44.4), with a difference of 2.31 standard errors ($elpd_{diff}$8.1, se 3.5). Thus, this proposed technique corroborates the conclusions of Part 1 and suggests that participants were relying on intra-item associations—associations that were distinct from the top-down influence of verbal representations of object identity.

The primary concern that prompted the introduction of a novel analysis technique was that the presently available methods did not account for the potential for trial-level correlations between the two measures to influence the conclusion drawn from the state

[10] See Appendix B for an overview of the WAIC and its use in model comparison

trace analysis. We pointed out that unmodeled negative correlation could bias a STA to falsely reject the unidimensional model. One advantage of the proposed technique is not only the removal of this bias, but also that the correlation can be estimated. In our data, it appears that there was a positive correlation between the two measures.

3.6 Remaining Shortcomings

One challenging limitation of this technique is computational. On a modern computer (24 GB Ram, 2.60 GHz) in which it was possible to run 6 Markov Chains in parallel, running the model on a simulated dataset of 8000 observations in which each individual chain had a 1000 sample warm-up and a 500-sample draw from the posterior required approximately 13 hours to finish. Fitting both the monotonic and non-monotonic model therefore required around 26 hours. In comparison, the technique of Kalish et al. (2016) required only a few seconds. Although the slow speed is acceptable when only one comparison between a monotonic and non-monotonic model is required, this makes it challenging to efficiently explore models from closely related families (e.g., some datasets may warrant interaction effects between condition and participants or condition and items). Likewise, designs that involve more than four conditions are expected to run even slower (given that more possible mixtures or orders may be possible). This latter example is unfortunate, given that additional conditions usually benefits a state-trace analysis by sampling more points along the state-trace in question.

A second shortcoming is that it will be important to better understand how this method handles participants with heterogeneous state-traces. That is, although the model estimates item- and subject-effects, it is not flexible enough to simultaneously model a group of participants that have both monotonic and non-monotonic state-traces (i.e., it

assumes that there is just a single, true ordering). In the context in which some participants exhibit a monotonic state-trace and other exhibit a non-monotonic state trace, it would be convenient to have a rough estimation on the power to detect those participants that exhibit evidence for the non-monotonic model. This unknown is intimately related to the first challenge, in that the full exploration of parameter space of possible orderings (i.e., 0% of participants exhibit a non-monotonic model up to 100% of participants exhibit non-monotonicity, in combination with number of participants and number of trials per condition) would require weeks to complete. This is currently under exploration.

3.6 Discussion

A state-trace analysis is a powerful, relatively non-parametric way to distinguish between models that differ in terms of latent dimensionality. However, despite its non-parametricity, the current approaches for implementing an STA are not assumption free. Specifically, current approaches assume functional independence between the two measures. It is shown here that this assumption of functional independence can be relaxed by exchanging it for the assumption that a hierarchical bivariate probit provides an adequate description of the data.

It is worth clarifying which parts of the STA technique introduced here are essential to the technique. The primary contribution of this technique is to demonstrate a method for instantiating uncertainty in the tonicity of condition effects while simultaneously modelling the correlation between measures that could take binary values. Uncertainty in the tonicity is instantiated as a mixture model of many possible orderings, and the correlations between measures are modelled by using a bivariate distribution.

Other details are nonessential. For example, the model was fit using Stan's No-U-Turn Sampler algorithm, but any suitably robust algorithm could have been applied. Similarly, model evaluation proceeded via comparison of WAIC. However, other measures might be more appropriate in other circumstances (Piironen & Vehtari, 2017; Vehtari et al., 2017). Parameters were modelled in a hierarchical setting to model variance due to item and subject effects, but the technique could be used with a shallower model. Finally, even the choice of a bivariate probit is nonessential. The bivariate probit distribution was chosen for its relationship to signal detection theory, but in principal other distributions that provide a reasonable description of the data could be utilized.

CHAPTER 4

VISUAL RECOLLECTION

4.1 Introduction

The primary question of this thesis is whether visual recollection can occur without mediation by episodic recollection. Part 1 presented an experiment that appeared to dissociate lateral, intra-item associations from vertical, visual part-verbal identity associations. In Part 2, a novel state-trace analysis was discussed that were able to reject the interpretation that the results of Part 1 were an incorrect by-product of unmodeled correlations between task performance. This final part utilizes the study paradigm of Part 1 and asks whether participants can utilize lateral associations to perform visual recollection.

A visual recollective account needs to be distinguished from a compound cue account. Compound cue accounts were initially proposed as an alternative to spreading-activation accounts of the influence of multiple cues on recognition tasks (Dosher & Rosedale, 1989, 1997; Ratcliff & McKoon, 1988). Briefly, a spreading-activation mechanism specifies that presentation of a prime results in the activation of representations that are related to the cues (e.g., via prior study). To the extent that a subsequently presented target is strongly associated with the cues, that target will have been "pre-activated" and processing of it can proceed either more quickly or more accurately. In contrast, a compound-cue mechanism functions without any pre-activation, but instead claims that multiple cues can interact to give a familiarity signal that is more informative than the sum of the familiarity signals from of the individual cues. The Part-Matching task from Part 1 in this thesis is an example of a recognition task involving

multiple cues, and the process of visual recollection would be analogous to a spreading activation account.

Two kinds of compound-cue mechanisms have been proposed. In the mechanism specified by Ratcliff & McKoon (1988), individual cues are combined super-additively, such that the familiarity signals from the cues combined by some function like multiplication (i.e., if the two items A and B are studied together and then A is used as a prime to the target B, participants' response for B will be based not only on their familiarity for A and their familiarity for B, but also the result of e.g. A*B). In a related account proposed by Dosher & Rosedale (1989), the component cues are stored in a composite (i.e., conjunctive) representation, and the enhancement in performance is through the additional familiarity signal conferred by this composite representation (i.e., if the three items A, B, and C are studied together and A are B as provided as cues at test, participants have a familiarity signal that is the familiarity of A + the familiarity of B + the familiarity of AB). The key to these compound-cue mechanisms is that multiple cues can interact in a super-additive way (e.g., either via a multiplicative operation performance on the cues or via storage of conjunctive whole) to provide a boost to familiarity which participants can use in a recognition task, and that this boost can occur without needing to propose that activation spreads to related targets.

To ask whether visual recollection can occur, an experiment was needed that can distinguish between a recollective versus a compound cue account. Although experiment 1 demonstrated that participants can learn intra-item associations, it does not speak directly as to how those associations were used during 2AFC. Under the visual recollection account, successful performance in the 2AFC would be contingent on

participants' ability to pattern-complete the remainder of the cued items. Pattern-completion on the matching pair would have produced (at least parts of) a single (correct) object, whereas pattern-completion of the mismatching pairs would have produced two conflicting images. Under the compound-cue account, the two matching cues would have simply interacted super-additively, and participants could have chosen the correct pair based on the higher familiarity signal induced by those matching parts.

To distinguish between a visual recollection and a compound-cue account requires a task in which participants must actively engage in pattern-completion in even the presence of a single cue (Figure 10). Experiment 2 achieved this with a speeded, perceptual detection task. Participants studied lists of images of objects, in which the objects were presented either binocularly or masked by CFS. At test, participants first engaged in the same cued-recall task as in Experiment 1 (Figure 10a). Then, immediately after the cued-recall, participants engaged in a go/no-go task (Figure 10b). In the perceptual detection task, the inverse of the cue, a "bullet-hole" stimulus appeared out of visual white noise. Participants were instructed to respond as soon as they were sure that an object was appearing from the noise. On some trials, no object appeared, and so participants' job was to wait the duration of the trial without responding. The object that appeared was always the bullet-hole that matched the cue.

Recollection and compound-cue accounts make contrasting predictions for the effect of a cue from a studied object on participants' response time (RT) distribution. Specifically, only a recollection account predicts that the leading-edge of the RT will shift earlier in time when the cue is from an object that was studied. This prediction of the recollection account follows from the assumption that, if the cue prompts a recollective

process, many of the visual details of the object that subsequently emerges out of the noise will have already been retrieved; the cue will enable a (perhaps implicit and incomplete) "preview" of upcoming the bullet-hole stimulus. The fastest response times on go trials should therefore be even faster when visual recollection has occurred, and the leading edge of the RT distribution would therefore be shifted earlier in time. In contrast, the compound-cue account would not predict an effect of study on the leading-edge. This lack of a leading-edge effect under the compound cue account follows from the assumption that the compound representation is not sufficiently activated by the single object part. That is, this prediction follows from the assumption that aperture parts do not provide a sufficient enough set of associates activate a substantial compound; for the compound cue to take effect and enable faster perception of the emerging object, the object must have already sufficiently emerged from the noise. Hence, the compound cue account would not predict that the fastest response times get faster with study.

The visual recollection account additionally predicts that the leading-edge effect should occur for objects that have been encoded primarily as lateral associations between intra-item features – encoded with minimal associations between the parts of the object and the identity of the object. That is, the extent to which the leading-edge effect occurs as a result of visual recollection as compared to episodic recollection is the extent to which the leading-edge effect occurs even when participants are unable to name the object when given the cue. To distinguish between a visual and episodic recollective mechanism, participants in this experiment studied objects either Binocularly or as masked by CFS. Given that the first experiment suggested that study of images of object that are masked by CFS preferentially enables participants to form intra-item associations

as compared to part-whole associations, the key result in this experiment was whether items that are studied under CFS and are not named produce a shift in the leading-edge of the RT distribution.

4.2 Methods

4.2.1 Participants

60 new participants were sought for this experiment. Four participants were excluded on the basis that they did not provide any responses to the cued-recall trials. An additional four participants were recruited to replace the excluded participants. In total 64 participants were run, but the data from only 60 are reported on. All participants were recruited from the University of Massachusetts undergraduate psychology program via SONA, and the local Amherst community. Participants received either course credit or monetary compensation (5$ minimum, with an additional 5$ for each 20 minutes of experimentation). Participants had either normal or corrected-to-normal vision. All participants provided written consent prior to participating in the experiment.

120 of the objects from experiment 1 were used. For each object, the cue in the cued-recall task was the same as in Experiment 1. Additionally, a single "bullet-hole" stimulus was constructed out of each object, which was the inverse of the cue. That is, the bullet-hole stimulus was the entire object image, excluding the circular aperture that was used as a cue. Objects were divided into 10 lists of 12 objects each.

Experiment 2 involved two additional calibration phases not present in Experiment 1. The first calibration phase was designed to test ocular dominance. In this phase, black arrows pointing to either the left or the right were presented on the screen.

The second phase was designed to calibrate the duration of time before participants would notice objects emerging from the noise. For this phase, an additional 32, gray-scaled images of unique, different objects were utilized. These objects were of the same size as were used in the primary experiment.

4.2.2 Procedure

4.2.2.1 Ocular Dominance

Ocular dominance for each participant was assessed using the technique introduced by Yang, Blake, & McDonald (2010). Across 100 trials, participants were asked to indicate the direction (either left or right) of an arrow presented centrally and masked by CFS. Each trial began with a fixation cross overlaid on the Mondrian mask was presented for a uniformly random duration between 400 to 600 ms. Following fixation, the arrow began to emerge out of the mask. On half of the trials (randomly intermixed), the arrow was presented to the left eye, and on the other half it was presented to the right. For each participant, direction of the arrow was counterbalanced with the eye it was presented to. The transparency of the arrow increased by 1% every 100 ms (in whichever eye it was being presented to).

Participants' dominant eye was chosen as the one for which, when the arrow was presented to this eye, participants' average response time was fastest. Objects that were masked by CFS in subsequent phases were only presented to the non-dominant eye.

4.2.2.2 Opacity Calibration

Prior to the main experiment, an adaptive procedure was used to calibrate the maximum opacity at which to present an object during CFS. Each trial began with

presentation of the CFS mask to both eyes. After a variable onset, the object was presented to the non-dominant eye at a 4% opacity (i.e., mostly transparent). The opacity of the object then increased linearly, up to a maximum value (defined below). The object remained at that maximum opacity until the final second of the trial, at which point it linearly decreased down to 0%. During these trials, participants were instructed to press 'Enter' on a keyboard to indicate when they first notice that an item has appeared.

After the presentation of each object, participants provided a PAS rating (Ramsøy & Overgaard, 2004). When participants provided a PAS rating of 3, they were instructed to respond more quickly, before the object had been identified. Similarly, when participants provided PAS ratings of 1, they were instructed to wait until they were sure that an item has appeared. To encourage participants to wait until they were sure that an item has appeared, 30% of all trials were "catch trials" in which no object emerged. Participants were given the same instructions on these catch trials as on the CFS trials: press 'Enter' when they first notice that an item has appeared. If participants mistakenly responded on catch trials, they received a warning that no object was present and reminded of the instructions that they should wait until they are sure that an item was appearing before responding. To encourage participants to respond as soon as they noticed an object, participants were informed of their reaction time by displaying it on the screen after participants provided the PAS rating —as well as whether an object was present.

The opacity required for 90% of the items to be noticed but not identified (i.e., to which a participant should provide a PAS rating of 2) after 3 seconds was estimated from these data.

4.2.2.3 Main Experiment

The main phase of Experiment 2 consisted of nine blocks each consisting of a study and test phase. During study phases, objects were presented for study under either CFS or Binocular viewing. Eight objects were studied in each study phase, four in each of the two study conditions. Presentation of the objects in the CFS condition proceeded as in the calibration phase of Experiment 1, with the maximum opacity continuously updated. Objects were presented twice, with the second presentation occurring only after all objects had been presented once. For the second presentation, the objects were presented in the same order as the first. As in the calibration phase, participants were instructed to indicate when they notice an object appearing during CFS trials (i.e., to self-terminate CFS trials when they could provide a PAS rating of 2, but no later), they were informed of their reaction time, and 30% of study phase trials were catch trials.

During test phases, participants encountered each object in the block on one test trial, in a randomized order. There were 12 items per test phase, the 8 items that were studied and 4 objects that were Not Studied. Conditions were counterbalanced across participants, but the same 12 items were always presented in the same list. On each test trial, participants were first required to perform the Naming (i.e., cued-recall) task from Experiment 1. They were encouraged to use their memory of objects they encountered from the study phase if they thought it would help. No feedback was provided. After participants provided a response, the aperture cue disappeared and was replaced by visual white noise. On $3/4$ of test trials (3 out of the 4 objects from each study condition, or 9 trials per list), after $400 - 600\ ms$ of white visual noise (randomly sampled from a uniform distribution), a matching bullet-hole stimulus gradually emerged out of the noise.

Participants were instructed to press a key as soon as they were sure that an object was emerging from the noise, but to not respond if no object emerged. The emergence was accomplished by linearly increasing the opacity of the bullet-hole stimulus, from 0% to 100% over four seconds, remaining at 100% opacity until either a participant responded or 10 seconds since the beginning of the emergence had passed. On the other fourth of test trials, no object emerged so participants were therefore required to wait without responding for 10 seconds. When participants responded while the object was emerging from the noise, they were shown their reaction time. When no object was present (i.e., if participants responded before the variable onset or on a no-go trial) they were warned that no object was presented and given a reminder that they should only respond when they are sure that an object was appearing.

4.3 Analyses

The primary questions were how the leading-edge of the RT distribution on go trials was influenced by study and whether a participant could correctly identify the object. Estimates of the minimum value of a distribution will be highly sensitive to noise. Rather than attempting to trim improbably fast responses (i.e., responses made in error before sufficient evidence had been gathered to make an informed decision – a designation that is difficult to make), two methods were chosen to analyze the leading-edge that would be relatively robust to noise.

First, a quintile analysis was performed on the RT distribution. A Linear Mixed Effects model was used that included a random intercept for both subjects and items, as well as fixed-effect intercepts (with all 2-way interactions and the 3-way interaction) for Condition (2 levels: Binocular and CFS), Percentile (5 levels: the quintiles of the RT

distribution), and Naming Accuracy (2 levels: correct and incorrect). Note that, by design, RTs in different quintiles are not independently distributed. However, the speed-up elicited by study can be assumed to be independent across quintiles. To account for this, this analysis was conducted on *differences in RT* as compared to the baseline of Not Studied. That is, the data entered into the model were not the percentiles from each of the three conditions but were instead the difference between the Not Studied Condition and either the Binocular or CFS condition. Analyzing differences maintains the independently distributed assumption of regression. The R packages lme4 (Bates et al., 2015) and emmeans (Lenth, Love, & Herve, 2018) were used for this analysis.

Although the quintile analysis provides a robust and relatively simple way to estimate the leading-edge, there are at least two ways in which the analysis could be augmented. First, the choice of percentiles to analyze is somewhat arbitrary. That is, there is little theoretical reason to prefer five percentiles instead of, e.g., four or six, each of which could also provide an estimate of the leading edge. Second, it would be beneficial for the estimate of the effect of one quintile to inform the estimate of the effect of the other quintiles. That is, by treating the quintiles simply as levels of a factor, the analysis does not leverage any of the wealth of information known about RT distributions.

The second analysis was designed to complement the quintile analysis. Here, the parameters of an exponentially-modified Gaussian (ex-Gaussian) distribution were fit to the data to summarize the key differences in the RT distributions. The ex-Gaussian is a three-parameter distribution, whose probability density function is given by the convolution of the probability density functions of an exponential and Gaussian distribution. One parameter corresponds to the scale parameter of the exponential, the

other two are the mean (i.e., location of the peak) and standard-deviation of the Gaussian component. Attempts have been made to link the parameters of the ex-Gaussian to latent cognitive processes, but the utility of such endeavors has been criticized (Matzke & Wagenmakers, 2009; Van Zandt, 2000). Still, the ex-Gaussian describes response time distributions well (Heathcote, Popiel, & Mewhort, 1991) and is used here to assess how the distributions differ across the conditions in order to test the *a priori* predictions, while remaining agnostic about whether the parameters correspond to particular cognitive processes. The parameters of the ex-Gaussian were fit in a hierarchical Bayesian framework. The peak location parameter of the Gaussian component was modelled with a random intercept for each participant and item, as well as an intercept and interaction for the study condition (3 levels) and intercept for whether the cue was correctly identified (2 levels). The scale and standard deviation parameters were allowed to vary according to whether participants correctly identified the object from the cue[11]. The model was fit using Stan and the brms package (Bürkner, 2017; Stan Development Team, 2017b).

4.4 Results

Participants largely followed instructions in both the study and test phases (Figure 11). During the study phases, participants provided predominantly PAS responses of either 2 or 3 during CFS trials (Figure 11a). Study improved cued-recall accuracy (Figure

[11] A few other, closely related models were considered. For example, models were considered in which the standard deviation or scale parameter was also allowed to vary by Condition, and with the interaction of Condition and whether the cue was correctly identified. Of the models considered, the one described in the main text fit the data best (as determined by WAIC). A comparison of all possible models – in which all possible permutations of condition, naming, participant, and item effects are considered for each of the three parameters – was not conducted given that the recommended use of WAIC is for adjudication between only a few, best models to avoid a selection induced bias (Vehtari et al., 2017).

11b)[12], largely replicating the trends of Experiment 1. Finally, go/no-go accuracy was almost perfect (Figure 11c); participants almost always responded when there was an object present and rarely responded when an object was absent.

The leading-edge of the RT distribution is the main observation of interest. Linear-mixed effects analysis revealed that the 20th percentile of response times for objects studied under CFS and not named were faster than those that were Not Studied and not named (Figure 12; compare leftmost orange datum to leftmost blue datum, in left panel).

In the fits of the ex-Gaussian distribution, four chains were used, each with 1000 warmup samples and 1250 samples from the posterior. All $split - \hat{R}$ were under 1.1 and there were no divergent transitions (Appendix A). Overall, the model appeared to have converged.

The main result of interest from this model is the posterior distribution for the effect of CFS-study on the mean component of the gaussian (as compared to Not Studied) when items were unnamed. That is, to what extent is the RT distribution shifted for items that were studied under CFS but not named? In the model, the 95% highest density interval of the difference in posteriors for the effect of CFS-study on the mean component of the gaussian from the effect of Not Studied when items were unnamed was below 0 (Figure 13). This corroborates the quintile analysis, suggesting that the entire RT

[12] Although cued-recall was overall lower in Experiment 2, there are at least two experimental design differences that may have lead to this. First, in Experiment 1 participants were asked questions about the studied objects to encourage learning. No such questions were asked during Experiment 2. Second, in Experiment 1, the 2AFC preceded the cued-recall. Although no feedback was provided after the 2AFC, participants effectively received additional cues in Experiment 1, as compared to Experiment 2.

distribution – including the leading-edge – was faster for items that were studied under CFS, even when those items were not named.

4.5 Discussion

In Experiment 2, it was asked whether recollection could proceed for non-episodic memories. Visual recollection was measured via a perceptual detection task that a followed cued-recall task. It was hypothesized that, if participants were able to engage in visual recollection, the aperture cue in the cued-recall task would have enabled retrieval of the visual details of the whole object. Those visual details would be the same details as subsequently emerge from the noise in the detection task. The extent to which participants' fastest RTs were faster after study would indicate the extent to which study enabled them to preview the bullet-hole stimulus. Importantly, for visual recollection to have occurred (as compared to episodic recollection), this leading-edge effect should have occurred even for objects that were studied under CFS, and even when participants could not retrieve the name of the object at study. These predictions of a visual recollection account were observed, suggesting that participants could engage in a recollection process for low-level, non-declarative representations.

Two features of the results are interesting to note. First, the leading-edge effect was large (Figure 12; compare the orange to either the blue or green lines in either panel). This was most likely brought about by the slow rate at which the bullet-hole stimulus emerged from the noise. The slow emergence of the stimulus was chosen to enable a larger window in which visual recollection could affect perception.

Second, although the earliest response times were faster for studied items that were both not-named and named, the shape of the RT distribution appeared to change

only for objects that were studied and named (Figure 12; compare left and right panels). That is, the data indicate that correct retrieval of the object's identity happened in concert with an additional process that affected RT in the perceptual detection task. One speculative account is that the change in shape is the result of recollection of features located further along the hierarchy of representations for object (i.e., episodic or "whole object" representations, rather than visual representations of object components). Under this proposal, the named cues lead to the generation of both low-level visual material, but also recurrent retrieval of visual details that were associated with the name of the object. The more associations that participants have learned (lateral and vertical), the more retrieval could compound. This effect would have been observed primarily on trials in which participants could name the cue because it was only on these trials that the recurrence had been initiated prior to perception of the object. That is, this speculation is largely just suggesting that the possible change in the shape of the RT that was observed for named cues (of objects that were studied) is a function of a larger amount of pattern completion (as compared to just visual recollection).

CHAPTER 5

GENERAL DISCUSSION

This thesis asked whether participants could recollect intra-item associations between the low-level visual features of objects that they were largely unaware of studying and were unable to identify at test. Part 1 presented Experiment 1, which dissociated the learning and retrieval of intra-item associations from part-to-object-label associations. The dissociation relied on a STA, however the currently available methods for conducting a STA ignore potential correlations between the dependent measures. In Part 2, a novel STA technique was introduced that explicitly models the correlations. Model comparison suggested that the results of experiment 1 were not a result of unmodeled dependencies between the two tasks. Together, Parts 1 and 2 established that associations between intra-item features can be learned separately from part-to-verbal-label associations. However, they could not distinguish between a visual recollection account of the memory retrieval process and a compound cue account. In Part 3, Experiment 2 was presented. In experiment 2, the speed of perception for missing intra-item associates following a cued-recall task was measured. RT distributional analyses revealed a shift in the leading edge of the RT distribution for objects that were studied with minimal awareness, even when those objects were not named. This shift in the leading edge was interpreted as indicating that the cued-recall task enabled participants to preview the upcoming bullet-hole stimulus before it emerged from the noise. That is, it the cued-recall task appears to have enabled participants to engage in visual recollection.

The RH account of memory predicted that recollection can support the production of arbitrary and novel associations between low-level associates. These behavioral

studies were therefore largely confirmatory hypothesis testing. However, progress in science ultimately requires falsification of theories. That is, understanding recollection and, more broadly, the relationship between mnemonic processes and representations, requires not only verifying that the predictions of one theory are not incorrect, but also that the predictions of competing theories are incorrect. Many theories that compete with the RH account do so at a neurocomputational level. At that level, the question is whether certain mnemonic processes are linked to particular brain regions. The RH account claims that mnemonic processes are not tied to specific brain regions (i.e., a process like recollection can occur throughout the MTL and the ventral visual stream). In contrast, "process-based" theories claim that particular mnemonic functions can only be performed by particular brain regions, e.g., that the hippocampus is responsible for recollection whereas the rhinal cortices are responsible for familiarity (Aggleton & Brown, 2006; McClelland, McNaughton, & O'Reilly, 1995; Montaldi & Mayes, 2010; Norman & O'Reilly, 2003; Squire & Dede, 2015; Squire & Zola-Morgan, 1991). Other "hybrid" theories exist between these two extremes, e.g., claiming that the hippocampus performs domain general associations whereas regions like the perirhinal cortex are primarily responsible for representing objects (Hannula, Libby, Yonelinas, & Ranganath, 2013; Wang, Yonelinas, & Ranganath, 2013). What do the behavioral studies in this thesis offer this neurocomputational debate?

The behavioral studies in this thesis complement prior neuroimaging experiments that have demonstrated that expectation can lead to the pre-activation of sensory information about the expected stimulus. Even when the cue is a relatively simple stimulus (e.g., textural fractal patterns, a tone), associated visual information (e.g., an

orientation) can be reinstated in cortex (Ekman, Kok, & de Lange, 2017; Hindy, Ng, & Turk-Browne, 2016; Kok, Failing, & de Lange, 2014). However, those studies have used associations across time (Ekman et al., 2017; Hindy et al., 2016), or across different sensory modalities (Kok et al., 2014; Kok, Mostert, & de Lange, 2017). Across many theories, this kind of highly associative information (across contexts or modalities) is often claimed to require the hippocampus (Clark & Maguire, 2016). However, the results of this thesis suggest that, even if these experiments could be replicated so as to not require such highly complex associations (e.g., only intra-object features), low-level cortical reinstatement may still be observable. Of the three kinds of accounts of memory, only the RH account clearly makes that prediction of a visual recollective process in low-level cortex.

However, Henke and colleagues have provided evidence that associative information can be learned implicitly (Henke, 2010; Reber & Henke, 2011; Reber, Luechinger, Boesiger, & Henke, 2012). In their neuroimaging work, the hippocampus appears to be critical for the formation of those associations. Indeed, it has been claimed that the hippocampus will always function to drive or initiate a pre-activation signal (Bosch, Jehee, Fernandez, & Doeller, 2014; Danker, Tompary, & Davachi, 2016; Hindy et al., 2016). Such a link between a cognitive process and an anatomical structure is inherently process-based and contends with the neurocomputational assumptions of the RH account. One study notes that many neuroimaging studies (and studies of recollection more generally) are often confounded by stimulus type, in that demonstrations of a link between recollection and the hippocampus tend to test only a certain class of stimuli – the representations of which likely require the involvement of the hippocampus (Ross, Sadil,

Wilson, & Cowell, 2017). After removing this confound, Ross et al. observed a recollective process that appeared to occur independently of the hippocampus. Indeed, at least one other recent theory posits that the involvement of the hippocampus – in a variety of cognitive processes – will be contingent on the representations required to perform the task at hand (Yonelinas, 2013). Elucidating the representations that are required for a given task and better understanding which anatomical structures are responsible for those representations will provide valuable data to further adjudicate between these competing accounts.

A brief note about the representations involved in this task, and those that are suppressed by CFS, is worth discussing. In the present experiments, intra-item parts were chosen subjectively. Similarly, CFS was chosen as a relatively coarse way to manipulate whether participants were able to learn higher-level representations. "Intra-item features" and "higher-level representations" are necessarily imprecise; it is not currently possible to quantitatively predict for an arbitrary visual stimulus which representations or associations participants have access to during study under CFS, nor which they utilize during visual recollection. However, some trends are taking shape in the CFS literature that the relationship of the visual properties of the mask and the masked image influence the efficacy of CFS. For example, CFS is more effective when the mask and masked image share spatial frequencies and cardinal orientations (Yang & Blake, 2012). Similarly, overlap between mid-level (i.e., representations that are not held in V1) prevents participants awareness of the masked images for longer durations (Cohen, Nakayama, Konkle, Stantić, & Alvarez, 2015). However, a systematic review of the comparison of how shared features between the masked and the masking stimuli

influence the duration of suppression is impeded the wide variety and non-systematic nature of the parameters used to generate the Mondrian mask (this is a recognized challenge and recent studies have given closer attention to such parameters, e.g. Drewes, Zhu, & Melcher, 2018; Zhu, Drewes, & Melcher, 2016). Moreover, the function relating how longer durations of suppression translate into the material that participants encode is not necessarily simple. As such, although the experiments in this thesis suggest that participants learned associations between, and pattern-completed intra-item features with minimal influence of episodic or declarative representations, conclusions should not be drawn about the specific intra-item features that were utilized in this study. Instead, the key result is that it appears that pattern-completion can proceed for relatively low-level visual representations, even with minimal mediation of higher level representations. This result fits most easily with an account that claims that mnemonic processes are not tied to specific representations.

CHAPTER 6

FIGURES

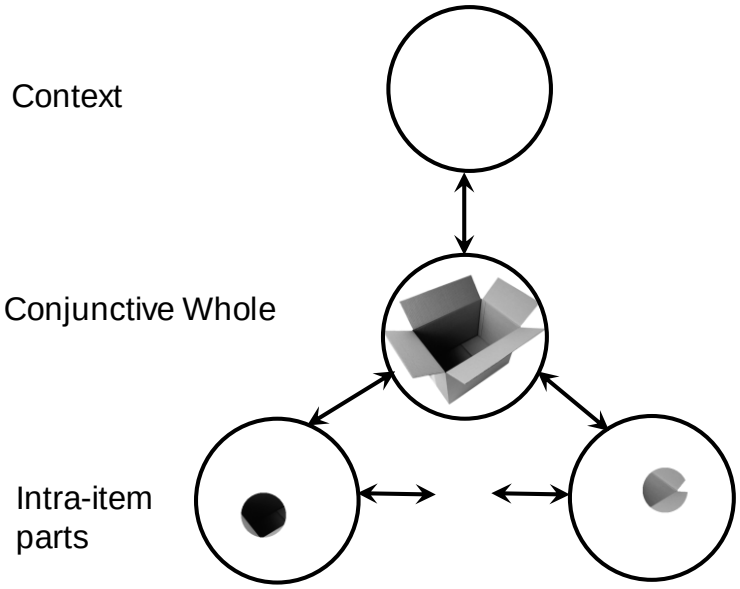

Figure 1 A subset of the representational hierarchy

Circles depict learned representations and arrows depict an association that has been learned between representations. Associations between representations within a level of the hierarchy (e.g., between intra-item parts) are termed lateral associations, whereas associations that span layers are termed vertical associations. Visual recollection is defined as the process whereby a cue of one intra-item part elicits retrieval of another intra-item part. Episodic recollection is defined as the retrieval of a higher-level representation (e.g., conjunctive whole, context). This association between intra-item parts is depicted as an incomplete arrow because it is unknown whether visual recollection can occur without mediation by vertical associations. Note that these figure is only designed as a schematic of the hierarchy; it is implicitly assumed in this thesis that there are many more levels above, below, and in between the levels depicted here.

Figure 2: Example stimuli used in the Gollin Incomplete Pictures task. Figure reproduced from Hirshman (1990).

In the Gollin Incomplete Pictures task, participants attempt to name line drawings of objects. The objects are presented one at a time in increasing stages of completeness (i.e. starting with the bottom image and progressing through the images from bottom to top). Participants perform this task across multiple sessions, sometimes across multiple days. Learning is indexed by the extent to which participants can identify the object at an earlier state of incompleteness across sessions.

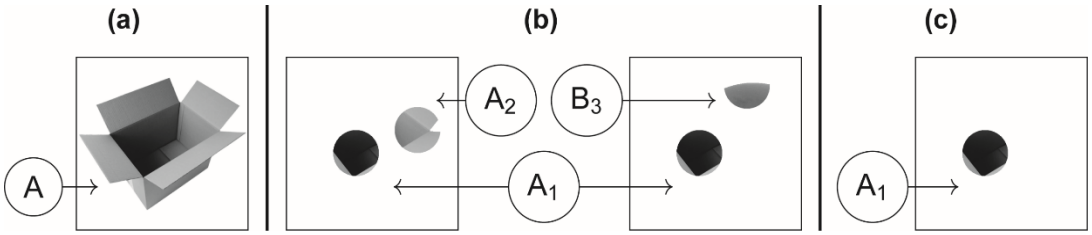

Figure 3 Sample test questions in Experiment 1.

(a) sample stimulus to study (object A). **(b)** sample 2AFC question between two matching parts both from the box (A_1 and A_2) and mismatching parts, one from the box (A_1) and one from a separate object (B_3). **(c)** sample naming question with the same 'box' stimulus. Labels identifying the origin of the parts are shown here for expository purposes and were not included in the experiment.

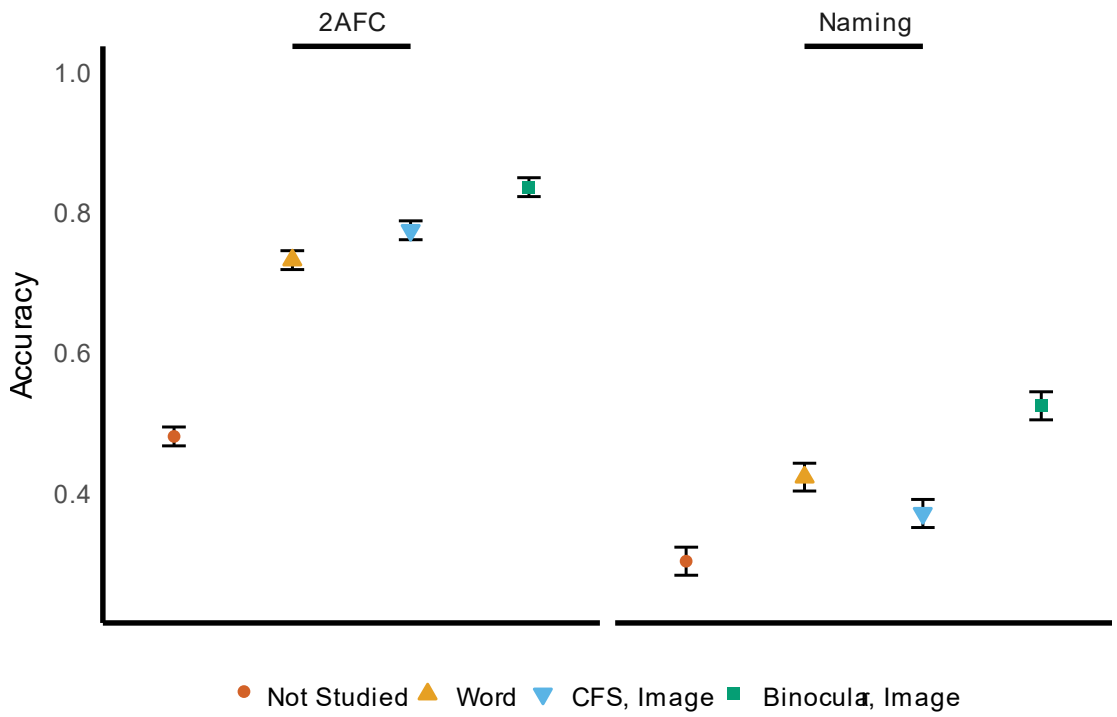

Figure 4 Crossover interaction was observed in Experiment 1.

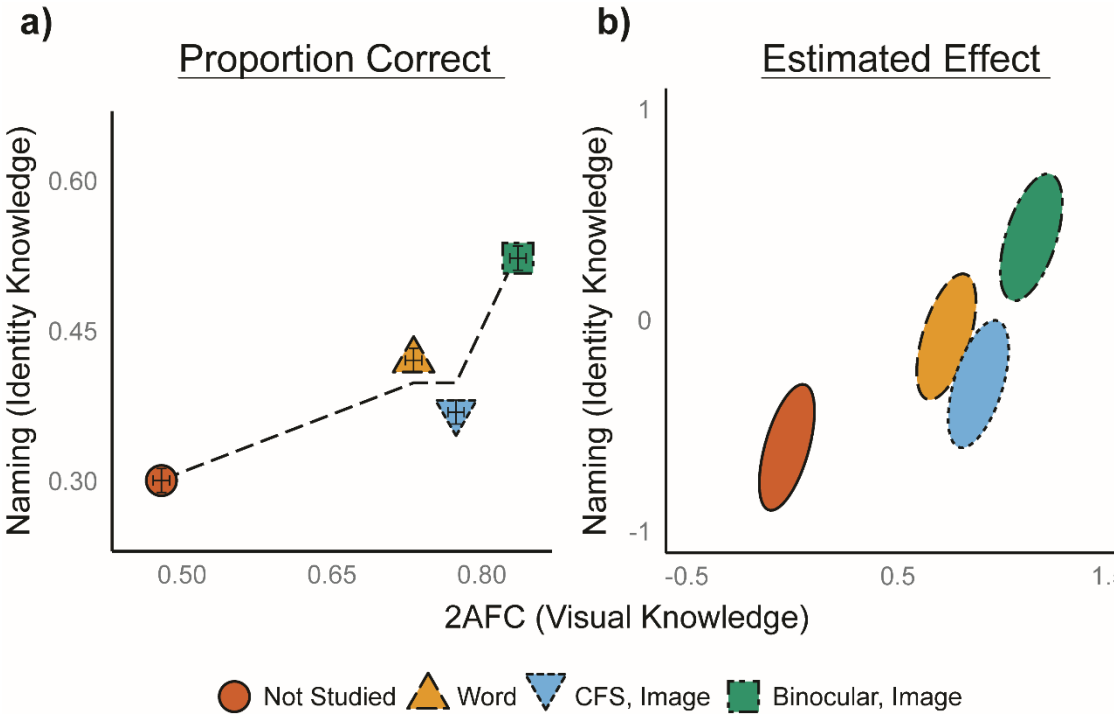

Figure 5 State-Trace analysis of Experiment 1

Average performance on the two tasks in the four conditions, with predictions of the best, single-latent-variable model overlaid as a dashed line; the fit of this single-latent-variable model was rejected in favor of a two-latent-variable model. Error bars indicate within-subject standard errors including correction by (Loftus & Masson, 1994).

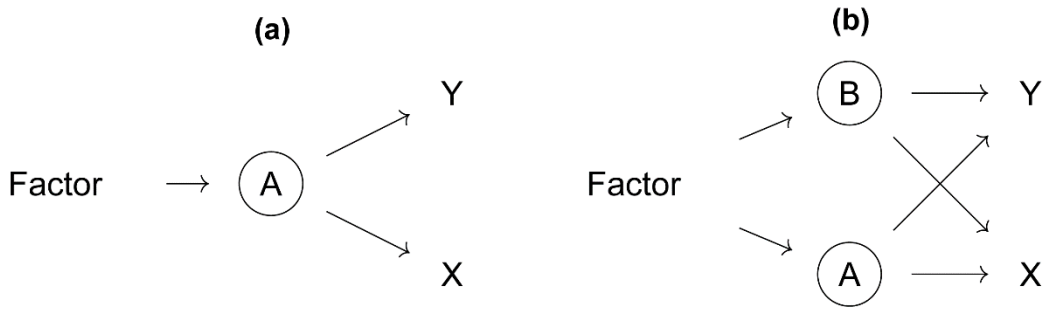

Figure 6. Hypothetical cognitive models comparable with state-trace analyses. Comparison of **(a)** unidimensional and **(b)** bidimensional cognitive systems. Reproduced from (Dunn, 2008). Factor indicates levels of an experimental condition. *A* and *B* refer to latent psychological variables of interest. *X* and *Y* refer to measured outcomes in an experiment.

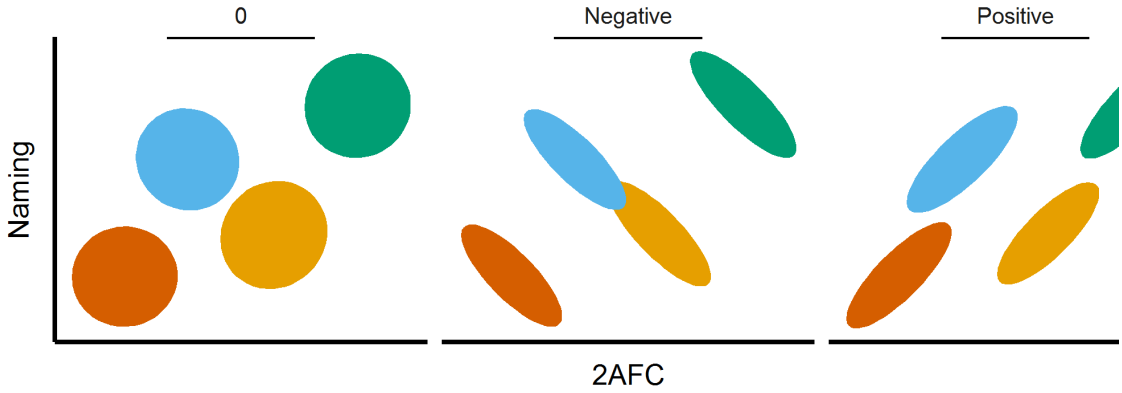

Figure 7. Negative effect of measurement dependence on STA.

Panels depict hypothetical data with 0, negative, and positive correlation between task performance. The method of Kalish et al. (2016) assumes the data look like the leftmost panel, but the middle and rightmost panel are also possible. Such dependencies could inflate or reduce the Type I error rate. For example, distributions that might appear to lie on a single state trace might more obviously lie on two when there is positive dependence.

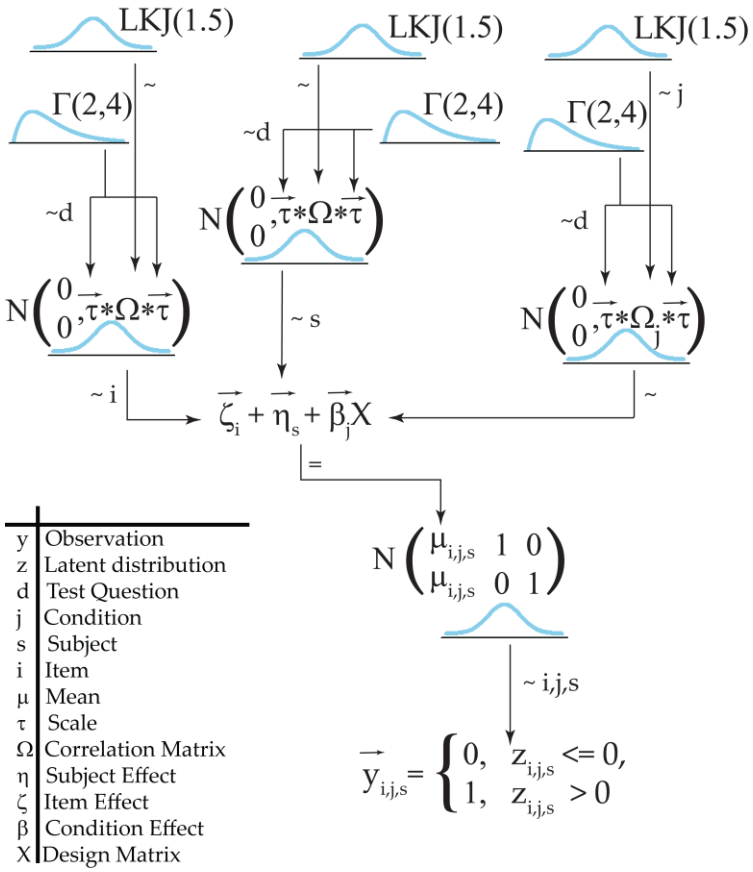

Figure 8. Schematic of the Bayesian hierarchical multivariate probit model. Each parameter of the model was sampled from a distribution, represented by a ~. A subscript indicates when a parameter was sampled for multiple levels of a factor (e.g., items, participants, or condition). Distributions are labeled with either *N* for normal (either univariate or bivariate), Γ for gamma, or *LKJ* for Lewandowski et al. (2009), which is a distribution of correlation matrices. In this diagram, vectors are indicated both by a column vector layout (as in the case of the means of the bivariate normal distributions) and via arrows. The diagram should be read beginning at the observation (y), upwards. Priors are given at the top of the diagram, indicated by numeric values. See text for further details.

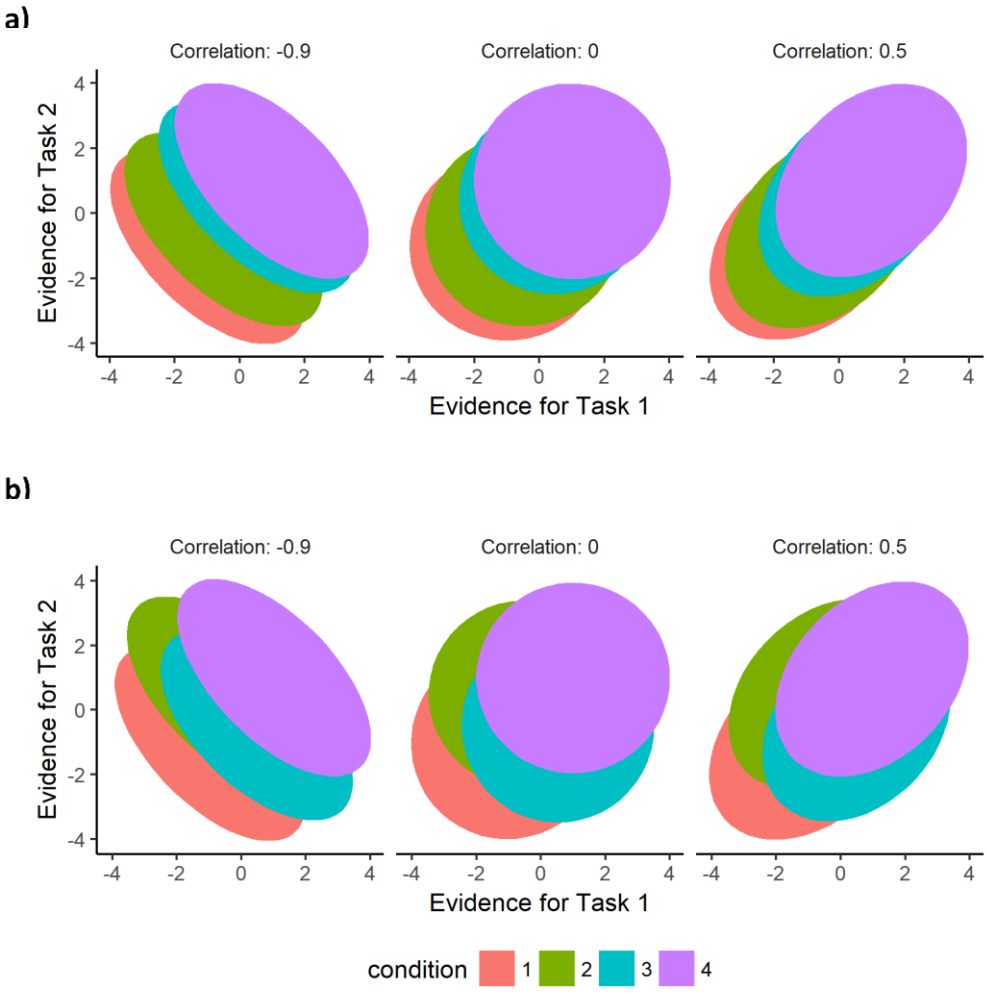

Figure 9 Simulated data for STA

a) Simulated monotonic and **b)** non-monotonic datasets, at each of 3 correlations. Data are plotted in the unconstrained space, implying values over 0 in either dimension would correspond to a correct response. Each dataset included 8000 observations. The ellipses indicate the 95% highest density interval of fitting a bivariate normal distribution to these simulated datasets. Notice that the distributions are highly overlapping.

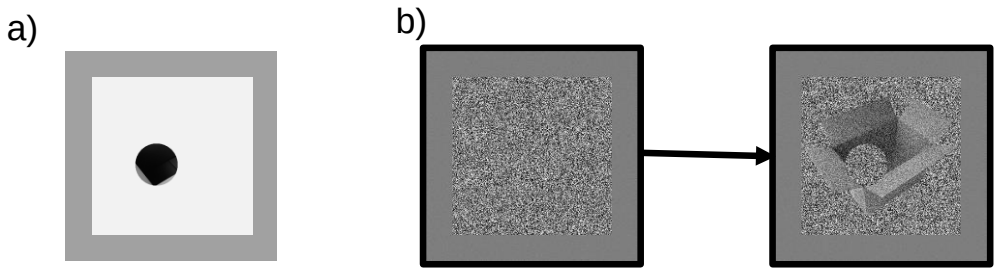

Figure 10 Experiment 2 task

a) Participants first completed a cued-recall, in which they encountered a circular aperture that was either studied binocularly, studied under CFS, or not studied. **b)** After the cued-recall, participants completed a speeded perceptual detection task in which the inverse of the aperture either appeared out of the visual noise, or nothing appeared out of the noise.

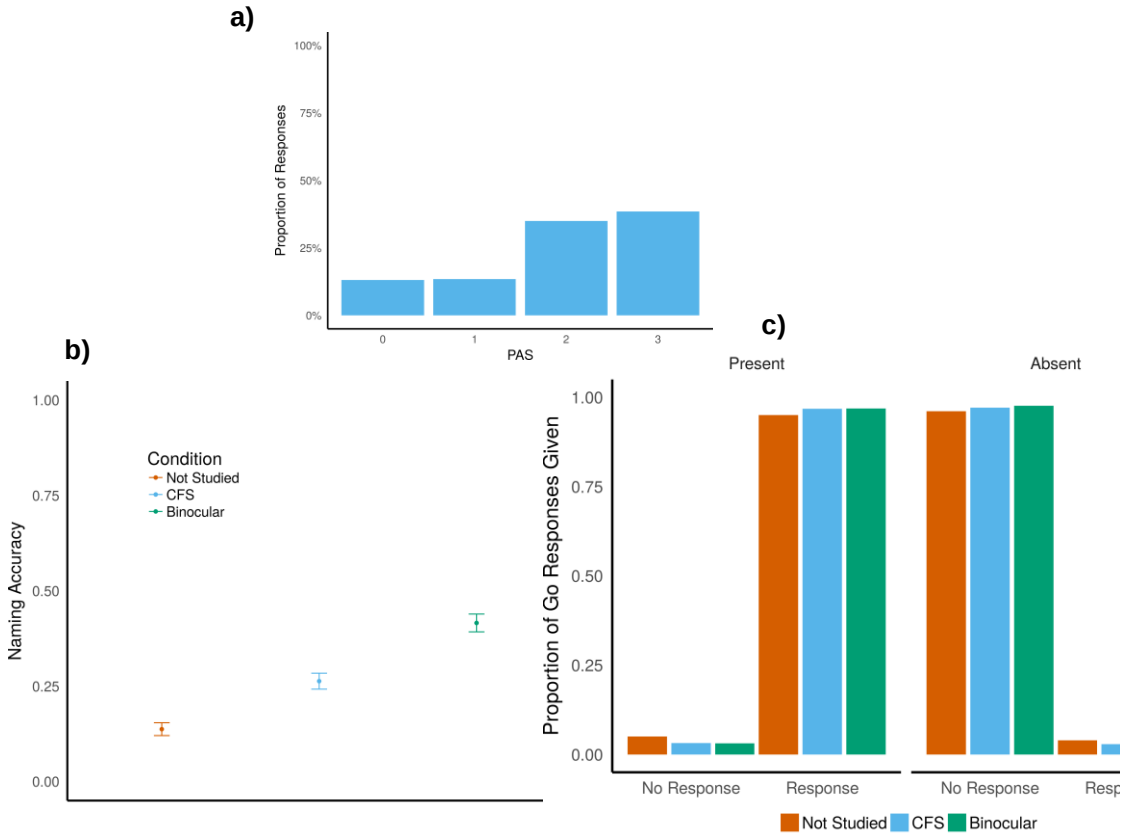

Figure 11 Descriptive plots of Experiment 2 performance

a) PAS ratings provided during the study phase. Participants followed the study instructions, providing predominantly PAS ratings of either 2 or 3. **b)** Naming accuracy on the cued-recall task. Naming accuracy increased with study. **c)** Go/no-go performance. Performance was nearly perfect, participants rarely responded when the target was absent and almost always responded when the target was present.

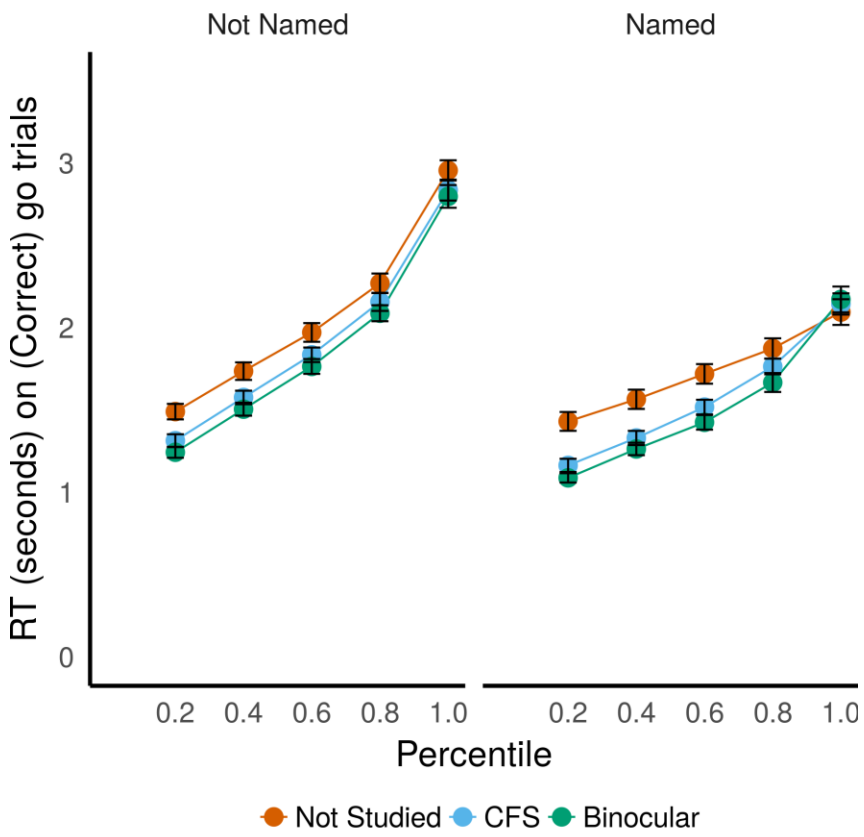

Figure 12 Quintile analysis of Experiment 2

RT quintiles in the go/no-go task are plotted as a function of whether the cue was named, and according to the study condition. Only trials in which participants correctly indicated that an object was present (correct go trials) are shown. Error bars indicate 95% Confidence Intervals around the mean.

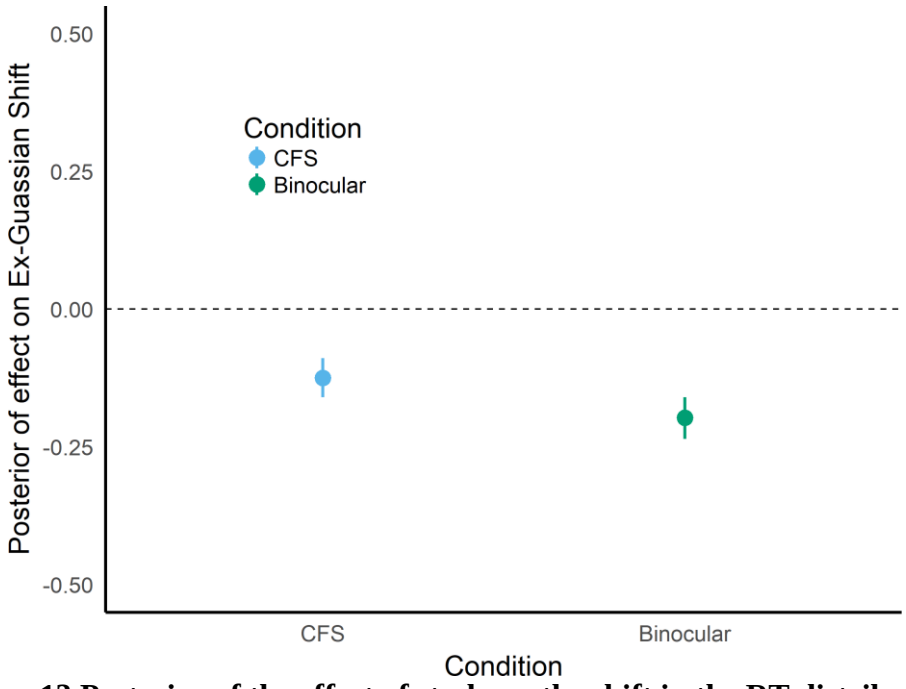

Figure 13 Posterior of the effect of study on the shift in the RT distribution. The difference in posterior distribution of the main-effect of study in the CFS and Binocular conditions, as compared to the Not-Studied baseline is shown. Values below 0 (dashed line) indicate a speed-up. The error-bars indicate the 95% Highest Density Interval.

APPENDIX A

METHODS TO ASSESS CONVERGENCE OF HAMILTONIAN MONTE-CARLO

A primary question of inference in a Bayesian setting is: what is the likelihood of a model given the data? Call the parameters of a model θ (which may be a single value or a vector of parameters, depending on the model) and data with n observations $y_1, \dots, y_n$ (where each observation may be a single value or a vector, as in the case of experiment 1 in which each observation was a pair of binary responses). To calculate that posterior probability $p(\theta|y)$, Bayes rule could be used

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p_t\,(y)},$$

where $p(\theta)$ is the prior probability of the parameters, $p(y|\theta)$ is the likelihood of the data given the parameters (referred to as the predictive density), and $p_t(y)$ is the likelihood of the data under the true data generating process. Unfortunately, Bayes rule is almost never applicable directly, given that $p_t(y)$ is unknown. Instead, an iterative algorithm that relies on the Monte-Carlo (MC) method is used. MC methods rely only on being able to calculate the relative likelihood of one set of parameters as compared to another, $p(\theta^1|y)\big/p(\theta^2|y)$. Given that $p_t(y)$ is common to both the denominator and numerator of that ratio, $p_t(y)$ cancels out and so the ratio is calculatable. A suitably chosen MC algorithm can therefore generate samples from the posterior distribution.

However, MC methods are only guaranteed to converge on the posterior distribution given infinite time, so some method of assessing their convergence is required in any applied setting. Two methods of convergence were used in this thesis.

One, the $split - \hat{R}$ is applicable to any MC method (Gelman & Rubin, 1992; Stan Development Team, 2016). The other, counting divergences, is unique to the Hamiltonian Monte-Carlo (HMC) method utilized by Stan.

The $\hat{R}$ (and its adaptation, the $split - \hat{R}$) relies on the intuition that, if multiple MC chains are constructed that start from a variety of starting points in parameter space, convergence will have been achieved if they are exploring similar regions of parameter space. It can be shown that an MC method will converge to a unique posterior distribution (e.g., Levin & Peres, 2017), so the extent to which the chains are generating similar samples is the extent to which they are generating from the same posterior distribution. The $\hat{R}$ is essentially an ANOVA calculated for each parameter, and its value can be interpreted by analogy with an $F - value$. An $\hat{R}$ of exactly 1 says that the variability of a parameter across chains is equal to the variability within chains; an $\hat{R}$ of 1 indicates convergence. A common heuristic states that an $\hat{R}$ greater than 1.1 indicates a lack of convergence (Stan Development Team, 2016).

The $split - \hat{R}$ is an updated $\hat{R}$ to account for the situation in which two chains are sampling from approximately the same region of parameter space, but they have non-random and nearly opposite trajectories through that space. Consider the case of two chains, each of which are generating samples from a single continuous parameter. One chain might start at a low value of that parameter and gradually produce larger and larger values of that parameter. The other chain might start at a high value of that parameter and produce gradually smaller and smaller values of that parameter. In that case, the within-chain and between-chain variance will be approximately equal (resulting in an $\hat{R}$ near 1), but the non-stationary behavior of the two chains indicates a lack of convergence. The

$split - \hat{R}$ accommodates this scenario by splitting each chain in half, calculating $\hat{R}$ on these half chains.

The number of divergences provides another check that the MC algorithm has not failed to converge, but this one is uniquely available to HMC (a type of which is implemented in Stan called the No-U-Turn Sampler) (Betancourt, 2017; Betancourt & Girolami, 2015). Divergences occur in regions of parameter space with rapidly changing predictive density. The HMC algorithm generates samples by moving along the gradient of the posterior distribution. Utilizing the gradient confers efficiency, enabling the algorithm can rapidly explore the full posterior distribution. However, when there are rapid and steep changes in the gradient, the chain may diverge, meaning that it proposes parameters that have infinitely miniscule likelihoods. Due to the difficulty of sampling from such regions, divergences (even just 1 or 2), indicate that the posterior distribution may not have been fully explored[13]. To properly sample from such regions requires either taking very small steps through parameter space (which is inefficient, will be much slower to converge, and if the steps are small enough loses the benefit of being able to notice regions where the parameter space is hard to explore), or by reparametrizing the model.

[13] Most other popular MC algorithm (i.e., Metropolis-Hastings algorithm or Gibbs Sampling), often miss such regions of extreme curvature due to their relatively inefficient exploration of parameter space. That said, if the region were discovered by such an algorithm, the lack of convergence could result in a $split - \hat{R}$ much larger than 1 if a chain gets stuck near a highly curved region.

APPENDIX B

WIDELY APPLICABLE INFORMATION CRITERION

The Widely Applicable Information Criterion (WAIC, also known as the Watanabe-Akaike Information Criterion) was used for model selection (Watanabe, 2010). The primary goal of this appendix is to provide a summary of the WAIC, validating it's use in model selection. The appendix is largely a summary of the description of the WAIC provided by Vehtari et al. (2017). A general overview of the use of predictive accuracy is presented next, followed by specifics of the WAIC.

Although Bayesian inference is not generally well suited or designed to make binary decisions about parameters in a model (i.e., concluding that an effect is 'significant' if a single parameter crosses some threshold), a useful way to express preference for one model as compared to others is to recast the inference as a question of predictive accuracy. That is, a model can be chosen based on whether it is 'best' able to predict new data in a replication. If a model makes highly precise predictions that turn out to be accurate in a replication, that model could be chosen as the model (among the models under consideration) that most closely resembles truth. In circumstances where a replication is not available, the ability to predict new data must be estimated based on the model's ability to capture the trends in the data at hand. However, the ability to capture all the patterns in the current data is often only conferred by extreme flexibility in a model. Often, that flexibility results in overfitting, meaning that an overly flexible model attempts becomes constrained by the idiosyncrasies of the data at hand (e.g., the performance of a particular participant in a particular trial). When overfitting occurs, the ability of a model to account for the data at hand will overestimate the ability of that

model to predict a replication. For this reason, most measure of predictive accuracy balance the ability of a model to predict the data at hand with some measure of the flexibility of a model.

The WAIC is a measure of predictive accuracy that balances the fit of a model to the present data with the variability in the number of ways that the model could account for the data. Using the notation from Appendix A, the predictive accuracy of a model in a replication, $\tilde{y}_1, \ldots \tilde{y}_n$, can be quantified with the posterior predictive distribution $p(\tilde{y}|y) = \int_\theta p(\tilde{y}_i|\theta)p(\theta|y)d\theta$. The posterior predictive distribution can be interpreted as, given the inferences that were made on the data at hand (i.e., which parameters, θ, were deemed most likely after observed y), what is the likelihood of the data that were observed in a replication? To have a single measure of predictive accuracy for models that account for datasets of different sizes and different numbers of parameters, the posterior predictive distribution can is replaced with the *expected log*[14] *pointwise predictive density (elpd)*

$$elpd = \sum_{i=1}^{n} \int p_t(\tilde{y}_i)p(\tilde{y}_i|y)d\tilde{y}_i.$$

The *elpd* for a new dataset can be interpreted as the amount by which a given datum was predicted ($p(\tilde{y}_i|y)$), weighted by the true probability of having observed each datum ($p_t(\tilde{y}_i)$), summed over the entire dataset.

[14] The logarithm is used mostly for computational reasons. Given that the probability of a dataset is the multiplication of the probability of each datum, the raw probability values would be miniscule – often impossible to represent on the precision afforded by a computer. Working on a log-scale keeps the probabilities in a manageable range.

Since the *elpd* for a new dataset involves unknowable terms (i.e., involves the likelihood of the true data generating process, $p_t(\tilde{y}_i)$), its value must be estimated. The fit of the current model is used to estimate the elpd via a term called the *log pointwise predictive density (lpd)* of the current dataset

$$lpd = \sum_{i=1}^{n} \log \mathrm{p}(y_i \,|y) = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|y)d\theta.$$

In practice, the *lpd* of the current dataset is computed using the results of an MC simulation. Calling $p_{post}(\theta)$ the posterior draws from an MC chain of length S (i.e., if θ is the vector of parameters produced at once step of the MC algorithm, each element of the chain is called $\theta^s, s = 1, \ldots, S$), define the *computed log pointwise predictive density ($\widehat{lpd}$)*

$$\widehat{lpd} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right).$$

The $\widehat{lpd}$ is the average probability of each datum in a dataset given the posterior distribution ($\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)$), transformed via a log, and then summed over each datum in the dataset.

Given the issue of overfitting mentioned above, the *lpd* and $\widehat{lpd}$ for the current dataset is an overestimate of the *elpd* for a new dataset. A penalty is required to account for differently flexibly models. Many of the measures of information criterion (e.g., WAIC, BIC, AIC, DIC, *etc.*) differ in the way in which they measure flexibility. Some methods simply count the number of parameters in a model. However, especially in the case of a hierarchical model, the number of parameters is a poor proxy for model flexibility. What is instead needed is some measure of the effective number of

parameters, p. So, for example, we can define the WAIC estimate of the computed *elpd* as

$$\widehat{elpd}_{WAIC} = \widehat{lpd} - \hat{p}_{WAIC}$$

The WAIC penalizes models based on the variability of the likelihood of the data.

$$p_{WAIC} = \sum_{i=1}^{n} var_{post}(\log p\,(y_i|\theta))$$

Again, this term involves an unknowable element (the variability of the log of the pointwise predictive density) and so must be estimated. The sample variance of a sample a of length S, defined as $V_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a}_s)^2$ (where $\bar{a}_s$ is the sample mean) provides the necessary estimation

$$\hat{p}_{WAIC} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\theta^s)).$$

The intuition behind using the variance of the pointwise predictive density is as follows (Gelman, Hwang, & Vehtari, 2014). Consider a dataset $y_1, y_2, \ldots y_n$, where n is small and the true data generating distribution is $y \sim Normal(0,1)$. A class of models for this data might be $y \sim Normal(\mu, 1)$, with a prior of $\mu \sim Normal(0, \sigma)$. Models with larger values of σ in the prior will be better able to capture more trends in y. When σ is very large, the model could easily account for a variety of types of data; the posterior distribution of μ will still put relatively large amounts of probability on a wide range of values for μ (given that the small dataset does not provide sufficient constraint on μ). That is, the variability of μ^s, for samples $s = 1,2, \ldots S$ will be high, and so will $\hat{p}_{WAIC}$. However, if σ were set to a small value (e.g., 1), the model is much less flexible, and the resulting posterior distribution of μ would be much narrower. A narrower posterior would

result in less variability in the likelihood of the data across each sample. $\hat{p}_{WAIC}$ will be, appropriately, much lower in this latter case.

The final two requirements of using a measure of predictive accuracy to decide between two models are 1) a measure of the variability in predictive accuracy, and 2) some way to determine when a difference in accuracy is 'significantly different.' To measure the expected variability in the WAIC, the calculation of standard errors proposed by Vehtari et al., (2017) was used. That is, for an individual model, $se\left(\widehat{elpd}_{WAIC}\right) = \sqrt{nV_{i=1}^{n}\ \widehat{elpd}_{WAIC,i}}$. The standard error of the difference in fit between two models (e.g., model A and model B) is given by

$$se\left(\widehat{elpd}^{A}{}_{WAIC} - \widehat{elpd}^{B}{}_{WAIC}\right) = \sqrt{nV_{i=1}^{n}\ \left(\widehat{elpd}^{A}{}_{WAIC,i} - \widehat{elpd}^{B}{}_{WAIC,i}\right)}$$

As a measure of the significance of the difference in fit, the threshold common in frequentist settings of approximately 2 standard errors was adopted. That is, models are considered to provide significantly different fits to the data when

$$\widehat{elpd}^{A}{}_{WAIC} - \widehat{elpd}^{B}{}_{WAIC} > 2se\left(\widehat{elpd}^{A}{}_{WAIC} - \widehat{elpd}^{B}{}_{WAIC}\right).$$

REFERENCES

Aggleton, J. P., & Brown, M. W. (2006). Interleaving brain systems for episodic and recognition memory. *Trends in Cognitive Sciences*, *10*(10), 455–463. https://doi.org/10.1016/j.tics.2006.08.003

Ashby, F. G. (2014). Is state-trace analysis an appropriate tool for assessing the number of cognitive systems? *Psychonomic Bulletin & Review*, *21*(4), 935–946. https://doi.org/10.3758/s13423-013-0578-x

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*(2), 137–181. https://doi.org/10.1016/0022-2496(79)90016-6

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. Retrieved from http://arxiv.org/abs/1701.02434

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In S. K. Upadhyay, U. Singh, D. K. Dey, & A. Loganathan (Eds.), *Current Trends in Bayesian Methodology with Applications*. CRC Press. Retrieved from http://arxiv.org/abs/1312.0906

Bosch, S. E., Jehee, J. F. M., Fernandez, G., & Doeller, C. F. (2014). Reinstatement of Associative Memories in Early Visual Cortex Is Signaled by the Hippocampus. *Journal of Neuroscience*, *34*(22), 7493–7500. https://doi.org/10.1523/JNEUROSCI.0805-14.2014

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436. https://doi.org/10.1163/156856897X00357

Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Clark, I. A., & Maguire, E. A. (2016). Remembering Preservation in Hippocampal Amnesia. *Annual Review of Psychology*, *67*(1), 51–82. https://doi.org/10.1146/annurev-psych-122414-033739

Cohen, M. A., Nakayama, K., Konkle, T., Stantić, M., & Alvarez, G. A. (2015). Visual Awareness Is Limited by the Representational Architecture of the Visual System. *Journal of Cognitive Neuroscience*, *27*(11), 2240–2252. https://doi.org/10.1162/jocn_a_00855

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2010). Components of recognition memory: Dissociable cognitive processes or just differences in representational complexity? *Hippocampus*, *20*(11), 1245–1262. https://doi.org/10.1002/hipo.20865

Danker, J. F., Tompary, A., & Davachi, L. (2016). Trial-by-Trial Hippocampal Encoding Activation Predicts the Fidelity of Cortical Reinstatement During Subsequent Retrieval. *Cerebral Cortex*, bhw146. https://doi.org/10.1093/cercor/bhw146

Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129. https://doi.org/10.1016/j.jmp.2015.08.004

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Dosher, B. A., & Rosedale, G. (1989). Integrated Retrieval Cues as a Mechanism for Priming in Retrieval From Memory. *Journal of Experimental Psychology: General*, *118*(2), 191–211. Retrieved from http://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=1989-31876-001&S=L&D=pdh&EbscoContent=dGJyMMvl7ESep7A4xNvgOLCmr1Cep7FSsam4Ta6WxWXS&ContentCustomer=dGJyMPPj33nmset55%2BS5febl8YwA

Dosher, B. A., & Rosedale, G. S. (1997). Configural Processing in Memory Retrieval: Multiple Cues and Ensemble Representations. *Cognitive Psychology*, *33*(3), 209–265. https://doi.org/10.1006/cogp.1997.0653

Drewes, J., Zhu, W., & Melcher, D. (2018). The edge of awareness: Mask spatial density, but not color, determines optimal temporal frequency for continuous flash suppression. *Journal of Vision*, *18*(1), 12. https://doi.org/10.1167/18.1.12

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446. https://doi.org/10.1037/0033-295X.115.2.426

Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, *47*(4), 389–416. https://doi.org/10.1016/S0022-2496(03)00049-X

Dunn, J. C., Kalish, M. L., & Newell, B. R. (2014). State-trace analysis can be an appropriate tool for assessing the number of cognitive systems: A reply to Ashby (2014). *Psychonomic Bulletin & Review*, *21*(4), 947–954. https://doi.org/10.3758/s13423-014-0637-y

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*(1), 91–101. https://doi.org/10.1037/0033-295X.95.1.91

Ekman, M., Kok, P., & de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, *8*, 15276. https://doi.org/10.1038/ncomms15276

Gayet, S., Van der Stigchel, S., & Paffen, C. L. E. (2014). Breaking continuous flash suppression: competing for consciousness on the pre-semantic battlefield. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00460

Gelbard-Sagiv, H., Faivre, N., Mudrik, L., & Koch, C. (2016). Low-level awareness accompanies "unconscious" high-level processing during continuous flash suppression. *Journal of Vision*, *16*(1), 3. https://doi.org/10.1167/16.1.3

Gelman, A. (2014). How do we choose our default methods? In *Past, Present, and Future of Statistical Science* (pp. 293–301). Chapman and Hall/CRC. https://doi.org/10.1201/b16720-31

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Graf, P., Squire, L. R., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(1), 164–178. https://doi.org/10.1037/0278-7393.10.1.164

Greene, W. H. (2017). *Econometric Analysis* (8th ed.). Pearson.

Greve, A., Donaldson, D. I., & Van Rossum, M. C. W. (2010). A single-trace dual-process model of episodic memory: A novel computational account of familiarity and recollection. *Hippocampus*, *20*(2), 235–251. https://doi.org/10.1002/hipo.20606

Hannula, D. E., Libby, L. A., Yonelinas, A. P., & Ranganath, C. (2013). Medial temporal lobe contributions to cued retrieval of items and contexts. *Neuropsychologia*, *51*(12), 2322–2332. https://doi.org/10.1016/j.neuropsychologia.2013.02.011

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience*, *11*(7), 523–532. https://doi.org/10.1038/nrn2850

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), 665–667. https://doi.org/10.1038/nn.4284

Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, *70*, 1–11. https://doi.org/10.1016/j.jmp.2015.10.004

Kent, B. A., Hvoslef-Eide, M., Saksida, L. M., & Bussey, T. J. (2016). The representational-hierarchical view of pattern separation: Not just hippocampus, not just space, not just memory? *Neurobiology of Learning and Memory*, *129*, 99–106. https://doi.org/10.1016/j.nlm.2016.01.006

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, S14. https://doi.org/10.1068/v070821

Kok, P., Failing, M. F., & de Lange, F. P. (2014). Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. *Journal of Cognitive Neuroscience*, *26*(7), 1546–1554. https://doi.org/10.1162/jocn

Kok, P., Mostert, P., & de Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, *114*(39), 10473–10478. https://doi.org/10.1073/pnas.1705652114

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition*. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Edition*. https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Lenth, R., Love, J., & Herve, M. (2018). Emmeans: Estimated Marginal Means, aka Least-Squares Means.

Levin, D. A., & Peres, Y. (2017). *Markov chains and mixing times* (Vol. 107). American Mathematical Soc.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319. https://doi.org/10.3758/BF03197461

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, *111*(4), 835–863. https://doi.org/10.1037/0033-295X.111.4.835

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252–271. https://doi.org/10.1037/0033-295X.87.3.252

MathWorks. (2015). MATLAB (R2015a). *The MathWorks Inc.* https://doi.org/10.1007/s10766-008-0082-5

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

Mayes, A. R., Holdstock, J. S., Isaac, C. L., Montaldi, D., Grigor, J., Gummer, A., … Norman, K. A. (2004). Associative recognition in a patient with selective hippocampal lesions and relatively normal item recognition. *Hippocampus*, *14*(6), 763–784. https://doi.org/10.1002/hipo.10211

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further Analysis Of The Hippocampal Amnesic Syndrome: 14-Year Follow-Up Study Of H.M. *Neuropsychologia*, *6*, 215–234.

Montaldi, D., & Mayes, A. R. (2010). The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. *Hippocampus*, *20*(11), 1291–1314. https://doi.org/10.1002/hipo.20853

Moors, P., Wagemans, J., van Ee, R., & De-Wit, L. (2016). No evidence for surface organization in Kanizsa configurations during continuous flash suppression. *Attention, Perception, & Psychophysics*, *78*(3), 902–914. https://doi.org/10.3758/s13414-015-1043-x

Murray, E. A., & Bussey, T. J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Sciences*. https://doi.org/10.1016/S1364-6613(99)01303-0

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*(8), 285–290. https://doi.org/10.1016/j.tics.2008.04.009

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, *20*(11), 1217–1227. https://doi.org/10.1002/hipo.20855

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203. https://doi.org/10.1037/a0033044

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1591–1607. https://doi.org/10.1037/a0028144

Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, *17*(1), 78–99. https://doi.org/10.1037/a0025809

R Development Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385–408. https://doi.org/10.1037/0033-295X.95.3.385

Reber, T. P., & Henke, K. (2011). Rapid Formation and Flexible Expression of Memories of Subliminal Word Pairs. *Frontiers in Psychology*, *2*, 1–11. https://doi.org/10.3389/fpsyg.2011.00343

Reber, T. P., Luechinger, R., Boesiger, P., & Henke, K. (2012). Unconscious Relational Inference Recruits the Hippocampus. *Journal of Neuroscience*, *32*(18), 6138–6148. https://doi.org/10.1523/JNEUROSCI.5639-11.2012

Ross, D. A., Sadil, P., Wilson, D. M., & Cowell, R. A. (2017). Hippocampal Engagement during Recall Depends on Memory Content. *Cerebral Cortex*, 1–14. https://doi.org/10.1093/cercor/bhx147

Roth, H. L. (2002). Effects of monocular viewing and eye dominance on spatial attention. *Brain*, *125*(9), 2023–2035. https://doi.org/10.1093/brain/awf210

Saksida, L. M., & Bussey, T. J. (1998). Toward a neural network model of visual object identification in primate inferotemporal cortex. In *Soc Neurosci Abstr* (Vol. 24, p. 1906).

Squire, L. R., & Dede, A. J. O. (2015). Conscious and Unconscious Memory Systems. *Cold Spring Harbor Perspectives in Biology*, *7*(3). https://doi.org/10.1101/cshperspect.a021667

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science*, *253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849

Stan Development Team. (2016). Stan Modeling Language User ' s Guide and Reference Manual.

Stan Development Team. (2017a). RStan: the R interface to Stan. Retrieved from http://mc-stan.org

Stan Development Team. (2017b). Stan Modeling Language Users Guide and Reference Manual. Retrieved from http://mc-stan.org

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), 1096–1101. https://doi.org/10.1167/4.8.61

van Tonder, G. J., & Ejima, Y. (2000). Bottom–Up Clues in Target Finding: Why a Dalmatian May Be Mistaken for an Elephant. *Perception*, *29*(2), 149–157. https://doi.org/10.1068/p2928

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*(3), 424–465. https://doi.org/10.3758/BF03214357

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*(1), 28–50. https://doi.org/10.1016/j.jmp.2003.11.004

Wang, W. C., Yonelinas, A. P., & Ranganath, C. (2013). Dissociable neural correlates of item and context retrieval in the medial temporal lobes. *Behavioural Brain Research*, *254*, 102–107. https://doi.org/10.1016/j.bbr.2013.05.029

Warrington, E. K., & Weixkrantz, L. (1968). New Method of Testing Long-term Retention with Special Reference to Amnesic Patients. *Nature*, *217*(5132), 972–974. https://doi.org/10.1038/217972a0

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Yang, E., & Blake, R. (2012). Deconstructing continuous flash suppression. *Journal of Vision*, *12*(3), 8–8. https://doi.org/10.1167/12.3.8

Yang, E., Blake, R., & McDonald, J. E. (2010). A New Interocular Suppression Technique for Measuring Sensory Eye Dominance. *Investigative Opthalmology & Visual Science*, *51*(1), 588–593. https://doi.org/10.1167/iovs.08-3076

Yonelinas, A. P. (1999). The contribution of recollection and familiarity contributions to recognition and source memory judgments: A formal dual-process model and an analysis of receiver operating charateristics.pdf. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*(6), 1415–1434.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yonelinas, A. P. (2013). The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behavioural Brain Research*, *254*, 34–44. https://doi.org/10.1016/j.bbr.2013.05.030

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800

Zhu, W., Drewes, J., & Melcher, D. (2016). Time for Awareness: The Influence of Temporal Properties of the Mask on Continuous Flash Suppression Effectiveness. *PloS One*, *11*(7), e0159206. https://doi.org/10.1371/journal.pone.0159206